THE COPAN TEAM
EDITED BY J.F. DONGES & J. HEITZIG

WORLD-EARTH DYNAMICS IN THE ANTHROPOCENE

A COPAN READER 2013–2021

POTSDAM INSTITUTE FOR CLIMATE IMPACT RESEARCH

Edited by J.F. Donges & J. Heitzig
Editorial assistance: J. Lochner (introductions & framing text)

doi:10.48485/pik.2021.004

published by potsdam institute for climate impact research
Member of the Leibniz Association
Telegrafenberg A 31, 14473 Potsdam, Germany

www.pik-potsdam.de/copan

First printing, July 2021

Contents

Thanks to John Schellnhuber for his inspiration

and to Wolfgang Lucht and Jürgen Kurths

for their invaluable support.

Introduction

copan coevolutionary pathways

PIK's 2013–2018 flagship project *copan — coevolutionary pathways*, since 2019 continued as the *copan* collaboration between PIK's FutureLabs on "Earth resilience in the Anthropocene" and "Game theory and networks of interacting agents", focusses on understanding and modelling the Anthropocene, the tightly intertwined social-environmental planetary system that humanity now inhabits. To this end, *copan* follows a social-ecological complex systems approach that allows to address the effects and limitations of human agency and system-level effects of networks and complex coevolutionary dynamics in the World-Earth system.

The project emerged from many informal discussions between leading PIK scientists, prominently involving Wolfgang Lucht and his ideas on a planetary social-ecology, and immensely influenced by PIK founding director John Schellnhuber's demand for "a genuine Earth System Analysis" in order help "secure an acceptable long-term coevolution of nature and civilization," which he described as "a cybernetic task for the emerging 'Global Subject'" [Schellnhuber, 1998]. It soon became clear that such an endeavor would from the beginning require a balanced mix between natural science based approaches (represented by Research Department 1) and complex systems science (represented by PIK Research Department 4) and would have to be very open to insights from and collaboration with the social sciences. In 2013, Jonathan Donges and Jobst Heitzig were assigned the task to explore such an agenda in a newly established flagship project, strongly supported by department heads Wolfgang Lucht and Jürgen Kurths. *copan* was named in reference to the ancient Maya city of Copán as just one example of a past civilization that displayed complex social-ecological dynamics leading to its eventual decline. The acronym officially stands for coevolutionary pathways, and the letter 'n' inofficially represents the heavy use of network-based modeling and analysis. Its logo combines a stylized form of the Maya language glyph 'Copán' and a visualization of coevolutionary pathways in the spirit of [Schellnhuber, 1998].

As of June 2021 the *copan* collaboration actively involves one intern, two bachelor, three master, five PhD students, one postdoc and five senior scientists. By the end of 2020, six Bachelor's, 25 Master's and five PhD theses have been completed. Furthermore, 67 papers were published in peer-reviewed journals. This work was done by current and former *copan* members, who are listed in the appendix of this reader. In addition, several software packages were developed as part of *copan* projects. To name the most important ones: the copan:CORE open World-Earth modelling framework and pyunicorn – Python modules for complex network and nonlinear time series analysis. A complete list can be found in the appendix.

This reader presents selected peer-reviewed and discussion papers put forward by *copan* at the Potsdam Institute for Climate Impact Research. The papers are logically ordered and sorted into topical sections on *Conceptual foundations and making the case* (Sect. 1), *Towards a unified analytical framework* (Sect. 2), *Theoretical and methodological work* (Sect. 3), and *Analyses and studies of concrete cases and contexts* (Sect. 4).

1
Conceptual foundations and making the case

This first section starts with two perspectives papers calling for a new way of looking at the Earth system and the human "World" build upon it.

In "Closing the loop: Reconnecting human dynamics to Earth system science" [Donges, J. F. and Winkelmann, R. et al., 2017], former and current PIK directors John Schellnhuber and Johan Rockström joined us to argue that the Anthropocene is dominated by planetary-scale social-ecological feedbacks and thus requires a new paradigm in Earth System science that is founded equally on a deep understanding of the biophysical and the social World-Earth System.

Adding to this, "The technosphere in Earth system analysis: A coevolutionary perspective" [Donges et al., 2017] stresses the importance of technological processes and proposes complex adaptive networks as a concept for describing the interplay of social agents with technospheric entities and their emergent dynamics for Earth system analysis.

In the Anthropocene, human actions have become critical to understanding planetary Earth system dynamics. To capture this in conceptual models, in the paper "Social tipping dynamics for stabilizing Earth's climate by 2050" [Otto et al., 2020a] we analyzed the importance of potential social tipping interventions for overall Earth system dynamics, using the terminology of tipping points in social-ecological systems as defined in the literature review [Milkoreit et al., 2018]. In addition, in "Human agency in the Anthropocene" [Otto et al., 2020b] we explored alternative concepts of human agency in the Earth system context.

The importance of this emerging perspective on the Earth system is emphasized in "Trajectories of the Earth system in the Anthropocene" [Steffen et al., 2018] with a small contribution of *copan* . This paper concluded that Earth system stewardship leading to transformative social-economic change is required to steer the Earth System away from risky "hothouse Earth" trajectories.

In order to get a first idea of what ingredients novel models of planetary-scale social-ecological coevolution might need to contain, we include here two quite different studies that feature selected social-ecological feedbacks.

The first exemplary modelling study, “Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution” [Barfuss et al., 2017] analyses the influence of heterogeneity on social interactions between the users of local renewable resources. We used a model of social learning agents in an adaptive network “copan:EXPLOIT” [Wiedermann et al., 2015], which will be introduced in more detail in a later section of this reader. Due to its simplicity, it has become a kind of paradigmatic example model used in several *copan* Master’s theses.

In the second exemplary modelling study, “Sustainability, collapse and oscillations in a simple World-Earth model” [Nitzbon et al., 2017], we use a very low-dimensional system of ordinary differential equations for modelling the coevolutionary dynamics of globally aggregated carbon, population, and capital stocks to demonstrate that the inclusion of socio-economic feedbacks can have a large influence on projected long-term Earth system trajectories.

THE ANTHROPOCENE REVIEW

Perspectives and controversies

The Anthropocene Review
2017, Vol. 4(2) 151–157

Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2053019617725537
journals.sagepub.com/home/anr
SAGE

Closing the loop: Reconnecting human dynamics to Earth System science

Jonathan F Donges,[1,2,*] Ricarda Winkelmann,[1,3,*] Wolfgang Lucht,[1,4] Sarah E Cornell,[2] James G Dyke,[5] Johan Rockström,[2] Jobst Heitzig[1] and Hans Joachim Schellnhuber[1,2]

Abstract

International commitment to the appropriately ambitious Paris climate agreement and the United Nations Sustainable Development Goals in 2015 has pulled into the limelight the urgent need for major scientific progress in understanding and modelling the Anthropocene, the tightly intertwined social-environmental planetary system that humanity now inhabits. The Anthropocene qualitatively differs from previous eras in Earth's history in three key characteristics: (1) There is planetary-scale human agency. (2) There are social and economic networks of teleconnections spanning the globe. (3) It is dominated by planetary-scale social-ecological feedbacks. Bolting together old concepts and methodologies cannot be an adequate approach to describing this new geological era. Instead, we need a new paradigm in Earth System science that is founded equally on a deep understanding of the physical and biological Earth System – and of the economic, social and cultural forces that are now an intrinsic part of it. It is time to close the loop and bring socially mediated dynamics explicitly into theory, analysis and models that let us study the whole Earth System.

Keywords

coevolutionary dynamics, complex adaptive networks, Earth System analysis, Earth System modelling, human agency, planetary boundaries, safe and just space for humanity, sustainable development goals

[1]Potsdam Institute for Climate Impact Research, Germany
[2]Stockholm Resilience Centre, Stockholm University, Sweden
[3]University of Potsdam, Germany
[4]Humboldt University Berlin, Germany
[5]University of Southampton, UK

*The first two authors contributed equally to this work.

Corresponding author:
Jonathan F Donges, Potsdam Institute for Climate Impact Research, Telegrafenberg A31, 14473 Potsdam, Germany and Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden.
Email: donges@pik-potsdam.de

Introduction

By pushing Earth's climate and biosphere out of the dynamics of the Holocene (Steffen et al., 2015a) humanity is at risk of moving our planet outside a safe operating space for humanity by altering important feedback loops, potentially producing abrupt and irreversible systemic changes with impacts on current and future generations (Steffen et al., 2015b).

From the start, Earth System science has recognized that humans are an important component of the contemporary system (Mooney et al., 2013; NASA, 1988). Integrating natural and social science perspectives on the Earth System has been a key aim of a suite of research initiatives over the past decades (e.g. AIMES, IHOPE, International Human Dimensions Program and Future Earth). Despite these efforts, key characteristics of the Anthropocene – human agency, global social and economic networks and important feedback interactions between human systems and planetary processes – have not been *dynamically* represented or otherwise resolved in existing Earth System and integrated assessment models.

Capturing these dynamics in a new generation of Earth System models should allow us to address a number of critical questions about socio-ecological turbulence in the Anthropocene, such as: Could transnational social movements such as the push for divestment from fossil fuels tip the socio-economics of carbon emissions? How is climate change science processed in world cultures and traditions other than those of the secular West? How are climate tipping events such as in the West Antarctic Ice Sheet interlinked with social and political transitions?

The biggest challenge in answering such questions is to understand human activities and social structures as the least predictable, but at present also the most influential component of our planet in the Anthropocene. This would, finally, contribute to closing the loop in theory, analysis and models of Earth System analysis (Future Earth, 2014; Schellnhuber, 1998, 1999).

To meet this challenge, Earth System analysis requires significant progress in three key areas forming the systemic substratum that many pressing, real-world sustainability questions have in common (Figure 1).

First: How best to represent human agency?

There is a long tradition of philosophical, anthropological, sociological and psychological research on the nature and degree of human agency, i.e. to what extent are humans free to act and what is the structure of the factors that constrain them. This has produced a wide variety of schools of thought, ranging from assumptions of substantial freedom of choice to behaviour within social norms and economic rules (Ajzen et al., 1991), to no agency at all (e.g. physics-based theories of social macrodynamics; Garrett, 2014, 2015). Here, we are primarily motivated to understand how this broad spectrum of (socially and structurally differentiated) human agency and behaviour can be appropriately included and evaluated in Earth System models. Our starting assumption is that we need to go substantially deeper than the common scenario approaches used in current Earth System modelling, where the dominant underlying social narrative is driven by macroeconomic optimization paradigms. These approaches, whilst computationally efficient, will necessarily exclude a wide spectrum of behaviours. Consequently, we call for new narratives of global change based on the fundamental *dynamics* following from different assumptions about human agency, and within such analysis for differentiation by social groups.

Second: What are the system-level effects of social networks?

The social is networked. Social interactions are mediated via information, trade, political and infrastructure networks. Such networks can change over time via adaptive, anticipatory and preference

Reproduced from: J. F. Donges et al., Closing the loop: Reconnecting human dynamics to Earth System science, Anthr. Rev., vol. 4, no. 2, pp. 151–157, 2017, doi: 10.1177/2053019617725537. Published under Creative Commons Attribution License 4.0 (CC BY NC).

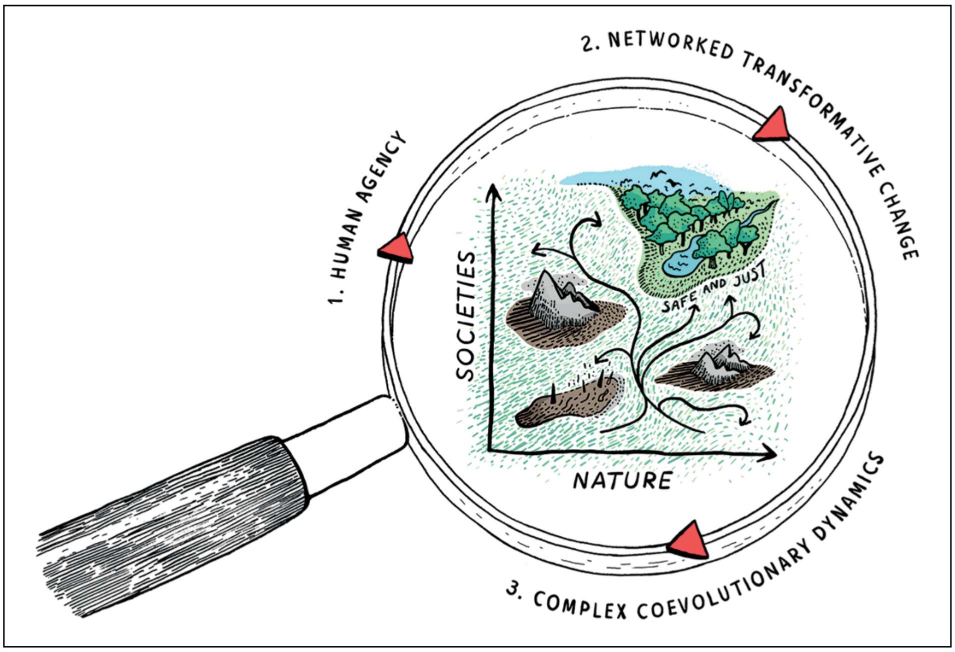

1. HUMAN AGENCY
2. NETWORKED TRANSFORMATIVE CHANGE
3. COMPLEX COEVOLUTIONARY DYNAMICS
SOCIETIES
NATURE
SAFE AND JUST

Figure 1. Closing the loop. Understanding and modelling the Anthropocene, the tightly intertwined social-environmental planetary system that humanity now inhabits, requires addressing human agency, system-level effects of networks and complex coevolutionary dynamics. The loop sheds light on a coevolutionary view of Earth System dynamics (Schellnhuber, 1998, 1999) in the Anthropocene including multiple development pathways, obstacles (mountains), dangerous domains (spikes) and the sought-after safe and just space for humanity (oasis).

formation processes. The dominant existing conceptualizations of Earth System loops – essentially using the same rigid box-and-arrow wiring diagram developed by the Bretherton Committee (NASA, 1988) – are no longer fit for purpose when the magnitude, direction of flows, and even composition of the components of the socio-environmental system are changing. Transformative phenomena such as the Great Acceleration (Steffen et al., 2015a) cannot be fully understood without digging into the network structure of the Anthropocene such as the wide-ranging teleconnections that emerge in land use change (Seto et al., 2012) and are the essence of digital communication between people. Earth System analysis needs to recognize that values and norms shape human behaviour, leading to changes in Earth System functioning with feedbacks to behaviours, values, and norms. This is a coevolving social-environmental network with an indisputably very rich structure.

Third: What tipping points and complex dynamics arise from social-environmental loops?

Even simple nonlinear systems can surprise us with our mostly linear thinking; even more so highly complex systems such as the Earth's climate. It is to be expected that social-environmental networks that feature myriad feedback loops will exhibit a wide range of complex behaviours. From observational records and modelling we know that there are several global-scale tipping elements in the climate system (Lenton et al., 2008; Schellnhuber et al., 2016). Even richer complex dynamics are expected and observed in the social sphere on comparably fast timescales (Bentley et al., 2014), particularly when interactions in the Anthropocene alter and strengthen feedbacks

Reproduced from: J. F. Donges et al., Closing the loop: Reconnecting human dynamics to Earth System science, Anthr. Rev., vol. 4, no. 2, pp. 151–157, 2017, doi: 10.1177/2053019617725537. Published under Creative Commons Attribution License 4.0 (CC BY NC).

between biogeophysical and social processes. Research and assessments ignoring the loops between and within these two spheres will inevitably overlook critical phenomena such as emerging multi-stabilities and tipping points. Models that allow for a systemic view that classifies potential pathways and identifies critical parameters, management options, windows of opportunity and dilemmas (Heitzig et al., 2016) represent important additions to studies more focused on quantification and prediction of individual trajectories.

A complex systems view of the Anthropocene

Effects that may arise even in simple systems due to complex dynamics may be illustrated for the case of a deliberately elementary representation of decarbonization in the energy sector. A dirty (CO_2-emitting) and a clean (e.g. sustainably renewable) energy technology compete while their market penetration can be influenced by a managing agent through subsidies. This is a hugely simplified case of the more general problem of multiple technologies, multiple economic incentive systems, non-economic values and, particularly, of a large number of interacting networked agents with different objectives and means. However, already this simple case system reveals non-trivial effects not usually taken into account in integrated Earth System modelling (Figure 2):

(i) A rich landscape of possible pathways exists that are sensitive to parameter settings and initial state. The cost-optimal pathway, an example of the imposition of a utility to be optimized (a very common practice in the analysis of such problems), is but one pathway toward a desired state and gives a rather incomplete picture of the dynamical landscape in which a manager is to operate. Closing the loop requires socio-ecological systems analysis. What is more, what is considered ‘desirable’ can differ among networked agents and potentially lead to conflict. Closing the loop means better inclusion of plurality of worldviews, priorities and objectives.

(ii) Large areas of parameter space form basins of attraction: pathways within these basins approach an end state that could have desired properties, but could also be an undesired state, underlining the importance for a manager to understand the structure of the dynamical landscape. Closing the loop means considering agency that is more multi-dimensional than single-purpose optimization, i.e. to follow broader concepts that allow potential access to a larger subset of trajectories.

(iii) Pathways toward a desired end state do not always initially lead in the direction of this state but can counterintuitively follow less obvious dynamical routes (which presents a problem to politics measured as short-term success). Along these lines, some paths that lead to desired end states have to temporarily traverse intermediate states with undesired properties (the situation must get worse before it will get better). Closing the loop requires a broader temporal perspective which may challenge short-term thinking in governance and policy making.

(iv) Pathways optimizing a given utility may display the phenomenon of ‘optimizing to the edge’, i.e. they tend to follow the edge of domains bordering undesired states, rendering them vulnerable against fluctuations that may tip them into neighbouring, less favourable domains of attraction. Closing the loop informs notions of desirability by explicit consideration of the resilience of trajectories.

This illustrative list of phenomena arising even in this simple example suggests that dilemmas in governing complex systems such as the global human–environment system (Heitzig et al., 2016)

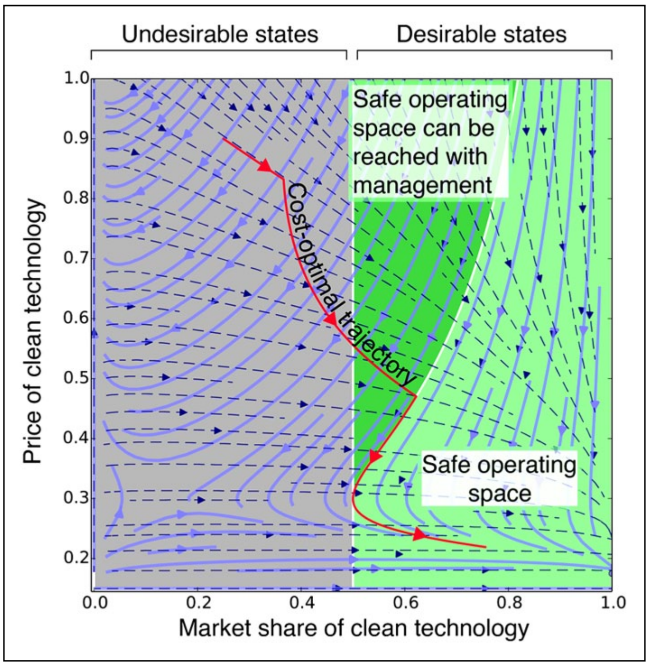


Figure 2. Complex dynamics arising from a conceptual model of decarbonization transformation. Mapping of trajectories in a dynamical system model of an energy market with competing dirty and clean technologies that can be influenced by subsidizing the clean technology (management). Business-as-usual trajectories without management (solid lines) as well as pathways with management (dashed lines) are shown. In this example, a market share of the clean technology larger than 50% is normatively considered as desirable. Background colours indicate state space regions such as the *safe operating space* (*shelter*, light green), where trajectories can remain in the desirable domain without management, or the region from which the *safe operating space* can only be reached through desirable states when applying subsidies (glade, dark green), following Heitzig et al. (2016). A typical cost-optimal pathway as generated by integrated assessment models is indicated by the red line.

require particular insight into three aspects of such dynamic landscapes. First, at issue is to what extent human intervention can alter the pathways upon which societies and the environment develop, i.e. what agency different types of agents have to manoeuvre on the landscape of trajectories, and what the instruments are to achieve this. Second, since humans act collectively as social groups on environmental processes and these are equally characterized by hierarchical interconnectedness, the macroscopic effects of coevolving complex networks on dynamic pathways have to be explored. And third, the topology of these dynamic landscapes has to be discovered as opposed to dissecting thin policy slices – this will require complex systems analysis, particularly regarding separation of domains of attraction, regions with steep gradients and faults, and critical dependence on key parameters.

Conclusion

We have shown how a simple model that explores trajectories towards decarbonization can produce complex behaviour and multiple outcomes, highlighting issues of agency over paths and of resulting complexity in the dynamical landscape of accessible paths. As such, this analysis demonstrates the utility of taking a complex systems, coevolutionary approach to dilemmas of the Anthropocene. This example highlights the first and third key area identified above. It is to be

Reproduced from: J. F. Donges et al., Closing the loop: Reconnecting human dynamics to Earth System science, Anthr. Rev., vol. 4, no. 2, pp. 151–157, 2017, doi: 10.1177/2053019617725537. Published under Creative Commons Attribution License 4.0 (CC BY NC).

expected that further complexities would arise by factoring in the collective effects of social networks on multiple agents and their interactions.

If science is to provide robust and useful input into this and other dilemmas that arise as a consequence of the transition to the Anthropocene, then Earth System models must embrace wherever possible these three areas: representation of socially differentiated agency, social-economic networks and complex coevolutionary dynamics. This would produce useful models of the Anthropocene (Donges et al., 2017; Verburg et al., 2016).

We see examples of such approaches emerging. For example, theory and models of biogeophysical dynamics in the Earth System are well established, and recently developed adaptive network approaches (Gross and Blasius, 2008) offer a flexible framework for modelling social-environmental regime shifts and transformations in an emergent and dynamic way without static prescription of scenarios, including phenomena such as social learning, segregation, norm and value change, and group dynamics such as coalition formation (Auer et al., 2015; Schleussner et al., 2016). Our vision for Earth System analysis calls for a synthesis of these so far disconnected phenomena within a complex systems framework.

The Paris climate targets (UNFCCC, 2015) and United Nations Sustainable Development Goals (UN SDGs, 2015) are examples of humanity's ambition to remain within a safe operating space at the same time as continuing to increase the wellbeing of the global population. Earth System science should play a critical part in this endeavour. To do so it must connect the behaviour and impacts of humans to biophysical processes and seek to understand the resulting very rich dynamics. We have existing tools and approaches to study such phenomena. Such analysis offers significant potential to augment existing models and methodologies and so help humanity chart a course towards a desirable Holocene-like Anthropocene.

Acknowledgements

We thank the participants of the LOOPS 2015 workshop on 'From Limits to Growth to Planetary Boundaries: Defining the safe and just space for humanity' in Southampton, UK, for fruitful discussions that substantially contributed to the ideas presented in this paper. This work was developed in the context of project COPAN on Coevolutionary Pathways at the Potsdam Institute for Climate Impact Research (www.pik-potsdam.de/copan). Karsten Bölts and Tim Kittel are acknowledged for support in creating Figure 2 and for their valuable contributions to the underlying research.

Funding

We are grateful to the Stordalen Foundation (via the Planetary Boundary Research Network PB.net) for supporting our research and the LOOPS workshop series. The Earth League's EarthDoc programme and the Leibniz Association (project DOMINOES) are acknowledged for further financial support in writing this paper.

References

Ajzen I (1991) The theory of planned behavior. *Organizational Behavior and Human Decision Processes* 50: 179–211.

Auer S, Heitzig J, Kornek U et al. (2015) The dynamics of coalition formation on complex networks. *Nature Scientific Reports* 5: 13,386.

Bentley RA, Maddison EJ, Ranner PH et al. (2014) Social tipping points and Earth systems dynamics. *Frontiers in Environmental Science* 2: 35.

Donges JF, Lucht W, Müller-Hansen F et al. (2017) The technosphere in Earth system analysis: A coevolutionary perspective. *The Anthropocene Review* 4(1): 23–33.

Future Earth (2014) *Future Earth Strategic Research Agenda*. Paris: International Council for Science (ICSU).

Garrett TJ (2014) Long-run evolution of the global economy: 1. Physical basis. *Earth's Future* 2(3): 127–151.

Garrett TJ (2015) Long-run evolution of the global economy: 2. Hindcasts of innovation and growth. *Earth System Dynamics* 6(1): 655–698.

Gross T and Blasius B (2008) Adaptive coevolutionary networks: A review. *Journal of The Royal Society Interface* 5(20): 259–271.

Heitzig J, Kittel T, Donges JF et al. (2016) Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system. *Earth System Dynamics* 7(1): 21–50.

Lenton TM, Held H, Kriegler E et al. (2008) Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences of the United States of America* 105(6): 1786–1793.

Mooney HA, Duraiappah A and Larigauderie A (2013) Evolution of natural and social science interactions in global change research programs. *Proceedings of the National Academy of Sciences of the United States of America* 110(Suppl. 1): 3665–3672.

NASA (1988) *Earth System Science: A Closer View*. Washington, DC: National Aeronautics and Space Administration.

Schellnhuber H-J (1998) Discourse: Earth System analysis – The scope of the challenge. In: Schellnhuber H-J and Wenzel DV (eds) *Earth System Analysis: Integrating Science for Sustainability*. Berlin/Heidelberg: Springer, pp. 3–195.

Schellnhuber H-J (1999) Earth system analysis and the second Copernican revolution. *Nature* 402: C19–C23.

Schellnhuber H-J, Rahmstorf S and Winkelmann R (2016) Why the right climate target was agreed in Paris. *Nature Climate Change* 6(7): 649–653.

Schleussner C-F, Donges JF, Engemann DA et al. (2016) Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure. *Nature Scientific Reports* 6: 30,790.

Seto KC, Reenberg A, Boone CG et al. (2012) Urban land teleconnections and sustainability. *Proceedings of the National Academy of Sciences* 109(20): 7687–7692.

Steffen W, Broadgate W, Deutsch L et al. (2015a) The trajectory of the Anthropocene: The Great Acceleration. *The Anthropocene Review* 2: 81–98.

Steffen W, Richardson K, Rockström J et al. (2015b) Planetary boundaries: Guiding human development on a changing planet. *Science* 347: 1,259,855.

UNFCCC (2015) Adoption of the Paris Agreement FCCC/CP/2015/L.9/Rev.1. Available at: http://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf.

UN SDGs (2015) United Nations Sustainable Development Goals. Available at: http://www.un.org/sustainabledevelopment/sustainable-development-goals/

Verburg PH, Dearing J, Dyke J et al. (2016) Methods and approaches to modelling the Anthropocene. *Global Environmental Change* 39: 328–340.

Reproduced from: J. F. Donges et al., Closing the loop: Reconnecting human dynamics to Earth System science, Anthr. Rev., vol. 4, no. 2, pp. 151–157, 2017, doi: 10.1177/2053019617725537. Published under Creative Commons Attribution License 4.0 (CC BY NC).

THE
ANTHROPOCENE
REVIEW

Perspectives and controversies

The Anthropocene Review
1–11
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2053019616676608
journals.sagepub.com/home/anr
SAGE

The technosphere in Earth System analysis: A coevolutionary perspective

Jonathan F Donges,[1,2] Wolfgang Lucht,[1,3] Finn Müller-Hansen[1,4] and Will Steffen[2,5]

Abstract
Earth System analysis is the study of the joint dynamics of biogeophysical, social and technological processes on our planet. To advance our understanding of possible future development pathways and identify management options for navigating to safe operating spaces while avoiding undesirable domains, computer models of the Earth System are developed and applied. These models hardly represent dynamical properties of technological processes despite their great planetary-scale influence on the biogeophysical components of the Earth System and the associated risks for human societies posed, e.g. by climatic change or novel entities. In this contribution, we reflect on the technosphere from the perspective of Earth System analysis with a threefold focus on agency, networks and complex coevolutionary dynamics. First, we argue that Haff's conception of the technosphere takes an extreme position in implying a strongly constrained human agency in the Earth System. Assuming that the technosphere develops according to dynamics largely independently of human intentions, Haff's perspective appears incompatible with a humanistic view that underlies the sustainability discourse at large and, more specifically, current frameworks such as UN sustainable development goals and the safe and just operating space for humanity. Second, as an alternative to Haff's static three-stratum picture, we propose complex adaptive networks as a concept for describing the interplay of social agents and technospheric entities and their emergent dynamics for Earth System analysis. Third, we argue that following a coevolutionary approach in conceptualising and modelling technospheric dynamics, also including the socio-cultural and biophysical spheres of the Earth System, could resolve the apparent conflict between the discourses on sustainability and the technosphere. Hence, this coevolutionary approach may point the way forward in modelling technological influences in the Earth System and may lead to a considerably deeper understanding of pathways to sustainable development in the future.

[1]Earth System Analysis, Potsdam Institute for Climate Impact Research, Germany
[2]Stockholm Resilience Centre, Stockholm University, Sweden
[3]Department of Geography, Humboldt University, Germany
[4]Department of Physics, Humboldt University, Germany
[5]The Australian National University, Australia

Corresponding author:
Jonathan F Donges, Earth System Analysis, Potsdam Institute for Climate Impact Research, Telegrafenberg A31, 14473 Potsdam, Germany and Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden.
Email: donges@pik-potsdam.de

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

Keywords
coevolution, complex adaptive networks, Earth System analysis, Earth System modelling, human agency, sustainable development, technosphere

Introduction

As a defining characteristic of the Anthropocene, human societies have created large-scale technological infrastructures such as world-spanning industrialized energy and food production and distribution systems for supporting historically unprecedented numbers of human beings embedded in increasingly complex socio-cultural structures while significantly intervening in the dynamics of the Earth System on a planetary scale. In this way, the worldwide evolving network of mutually interdependent technological and social macrostructures (examples for the latter include modern states, bureaucracies and social institutions in general), the *technosphere* in the sense of Haff (2014a), gives rise to key global environmental crises. These crises and their local and regional manifestations are reflected in the transgression of planetary boundaries such as those related to anthropogenic climate change, degenerative land-use change, accelerated biodiversity loss, perturbation of the global biogeochemical cycles of nitrogen and phosphorus, and the creation and release of *novel entities* such as nanoparticles and genetically engineered organisms (Steffen et al., 2015b).

In this contribution to the *Anthropocene Review*'s Special Issue on the technosphere, we reflect on the implications and relevance of Haff's concept in the context of Earth System analysis. This field of research explores possible future development pathways compatible with the coevolutionary dynamics of the biogeophysical and socio-technological spheres and aims at identifying management options for navigating to sustainable safe operating spaces while avoiding undesirable Earth System states such as 'catastrophe domains' (Schellnhuber, 1998, 1999). Our contribution intends to connect separate discourses about the technosphere on the one hand, and Earth System analysis and sustainable development on the other hand, by providing insights into current debates on how to include technological dynamics in Earth System models and exploring how the concept of the technosphere could be used to advance the understanding of these dynamics.We begin by discussing human agency in Haff's technosphere concept from the perspective of sustainability science. Then we briefly consider the relevant state-of-the-art of modelling technological dynamics in Earth System science and discuss issues of collective human agency in this context. Finally, we propose a complex systems approach for analytically dealing with the technosphere in the Earth System that is founded on (1) coevolutionary dynamics and emergence and (2) adaptive Earth System networks.

The technosphere and human agency

Agency is a key concept in the Anthropocene discourse. It arises as a crucial issue when considering an Earth System that is not only influenced by a socio-technological complex but also generates with increasing severity unintended consequences from the actions of that complex with repercussions for human societies. The notion of agency is traditionally debated in philosophy and sociology, but has received much attention as well in psychology and neuroscience in the last decades. Put simply, in these fields agency is the human experience of being the subject or owner of one's actions. This sense of agency is refined in the philosophy of action, where the term usually refers to the capability of an agent to perform deliberate and intentional action as opposed to forced, determined or random behaviour (Moya, 1990; Schlosser, 2015). The sociological concept of agency is often used as an antonym to social structure (Elder-Vass, 2010; Ritzer, 2010). On the one hand, structure

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

determines the individual's actions and behaviour. On the other hand, structure emerges from the actions of individuals, forming a coevolutionary loop (Snijders et al., 2010). The concept of agency of the individual emphasizes some degree of potential primacy of the individual over structure. Thus agency can be understood as one part of a dialectic understanding of the social.

Haff's concept of the technosphere shifts the focus from social relations to relations between humans and technology, a theme that is explored from other perspectives in the field of science and technology studies (e.g. Latour, 2014). Haff raises important questions regarding human agency and the controllability of large-scale technologies as well as the role of technology in the interrelation between human societies and other parts of the Earth System. Haff attempts to take a physicist's outside point of view on the technosphere as a 'geological phenomenon', postulating that the technosphere follows some 'physical law' or 'quasi-autonomous dynamics' such as the principle of maximum entropy production (Haff, 2014a). From this perspective, human agency and purpose may have been the originators of technological systems, but are no longer their controlling factor. Haff thus presents an account of recent human development as only a part of the systemic dynamics of the technosphere, thereby challenging the intuition that political decisions and societal change are solely the result of human volitions.

Haff notes that human actions are strongly constrained by technological possibilities and dependencies. Technologies and institutions increase societies' robustness to external and internal disturbances but also constitute path-dependencies and lock-ins that make large-scale changes difficult. The energy system is a good example of such a lock-in: Investments in fossil fuel technology can be considered as costs that owners of such investments wish to recover (and large parts of society wish to make use of). A radical shift in energy production towards renewable energies would make these prior investments worthless. Thus, a rapid transition to renewable energies is proving to be difficult.

Haff takes this argument to the extreme: Motivated by an apparent separation of scales between the level of the individual and the large-scale technological complexes, as suggested by Haff's rules of inaccessibility and impotence (Haff, 2014a), humans as individuals do not, in his view, exert direct influence on the dynamics of the technosphere and hence its repercussions (Haff, 2014b). Similarly, other authors argue that social metabolism can be described as a thermodynamic machine with intrinsic momentum originating in the flows of energy and material required to construct, maintain and transform large-scale infrastructures (Garrett, 2014, 2015). Haff puts this extreme position only partly into perspective, by focusing on leadership and control. Even if humans might have agency on an individual level, he argues that they do not have it on the aggregate level. Instead, the argument in Haff's papers suggests that the technosphere has non-human agency, which is in line with the discourse on the possibility of the emergence of general artificial intelligence and its consequences (Bostrom, 2014). The technosphere is presented as an emergent super-organism with its own teleology, desires and needs (Haff, 2014a), rather than serving human needs and normative goals.

Let us follow, for a moment, the assertion that the technosphere follows its own independent dynamics. This would imply that there is little room for political or ethical choice on a planetary level, e.g. for an intentional shift of technology towards sustainable production. Without the ability to influence technological development at a large scale, efforts to establish and implement normative goals such as UN sustainable development goals (Griggs et al., 2013) and frameworks such as sustainability paradigms (Schellnhuber, 1998, 1999) and the safe and just operating space for humanity (Raworth, 2012; Rockström et al., 2009) would be futile. Taking this into consideration, Haff's concept of the technosphere is incompatible with a sustainable development discourse founded on humanitarian principles.

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

However, we currently see at best mixed evidence for technology following its own dynamics in a manner that is totally independent of human intentions. Certainly, deviations from the business-as-usual path require more effort to succeed because well-established vested interests and technological inertia have to be overcome. But there is, in our view, no a priori reason why normative goals are unachievable; there are many examples where policies can opt out of large-scale technologies (e.g. the global banning of CFCs and the nuclear fade out in Germany). Instead of taking the development of technology as given, we suggest to do the opposite: frame it as a political question, i.e. regarding collective social action.

This is perhaps the most fundamental shortcoming of Haff's technosphere concept as it stands: the fact that humans reflect on their relationship with the world and adapt their actions accordingly does not seem to have any consequence for the emergent phenomenon of the technosphere. But history, for example of economic institutions, shows that theories about human societies and their environments can influence their behaviour, sometimes even leading to situations of self-fulfilling or self-defeating prophecy (e.g. Ferraro et al., 2005). Therefore, we think it is essential to consider human reflexivity as an integral part in the coevolution between technology and human societies.

In the following, we aim for a more differentiated understanding on the technosphere concept building on Haff's notion that

> The technosphere includes the world's large-scale energy and resource extraction systems, power generation and transmission systems, communication, transportation, financial and other networks, governments and bureaucracies, cities, factories, farms and myriad other 'built' systems, as well as all the parts of these systems, including computers, windows, tractors, office memos and humans. It also includes systems which traditionally we think of as social or human-dominated, such as religious institutions or NGOs. (Haff, 2014a)

While Haff capitalizes on a geophysical perspective on the dynamics of large-scale technological systems, he still includes social-dominated systems into his picture of the technosphere. To develop our arguments further, we attempt to distinguish more clearly between those two classes of phenomena that are emergent from the point of view of human individuals and technological objects: (1) macrosocial entities and structures such as social networks, governments and bureaucracies, religious institutions or non-governmental institutions (NGOs) and (2) technological macrostructures such as the internet or large-scale energy and resource extraction and transport systems. While such a classification is not always strictly feasible because of the myriad of interdependencies and co-enabling effects in densely entangled social and technological macrosystems, it is useful for distinguishing agency on the level of human individuals with respect to macrosocial entities and structures from the *macro-agency* of social macrostructures with respect to technological macrostructures. We refer to macro-agency as the *collective agency* of social macrostructures in the sense of their capability to govern, influence, direct and transform technological macrostructures. It should be stressed that this macro-agency arises from the individual agencies and is not an expression of an independent will, it is an emergent macro-phenomenon of networked individuals. Macro-agency differs qualitatively from the agency of human individuals because it is subject to distinct and strong path-dependencies and self-set rules.

Representation of technological systems in Earth System modelling

In Earth System analysis, mathematical and computer models are used as the main analytical tool to gain insights into the functioning and future development of components of the Earth System

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

and of the system as a whole. However, the representation of human societies and technology pose great challenges to formal modelling. Human activities as a whole are modelled in a number of different ways at several scales (Verburg et al., 2016). At present, most global models such as those employed in the assessment reports of the Intergovernmental Panel on Climate Change (IPCC) do not do an adequate job of simulating the human component of the Earth System in a dynamical way. Most global-level representations are based on general equilibrium models of the economy, which often do not include non-linear dynamics (e.g. feedbacks and emergent properties from agent interactions) and are based on strong assumptions about aggregate economic behaviour. For example, integrated assessment models typically only couple biophysical Earth System models (normally climate models) with economic models in a simple, one-way direction (van Vuuren et al., 2012). On the other hand, complex system approaches, such as agent-based models and simple conceptual (toy) models, generally do not operate at the large regional or global levels.

Perhaps an exception to this assessment, while still lacking representations of emergent social-technological structures and dynamics, is the World3 model, made famous by its use in the Limits to Growth scenarios (Meadows et al., 1972). The World3 model is basically a systems dynamics model that is organized around five sectors – population, capital, agriculture, non-renewable resources and persistent pollution (Costanza et al., 2007a). So although it does not contain an explicit technosphere module, World3 does simulate the metabolism of the technosphere – that is, the human commandeering of energy and resources and the expulsion of pollutants into the Earth System – and some of the critical feedbacks associated with this metabolism. Importantly, the model describes the metabolism of the technosphere as a deterministic dynamical system without invoking explicit representations of the agency of a social planner seeking optimal trajectories according to some prescribed utility function. Intriguingly, World3 does a remarkably good job of simulating the observed metabolism of the technosphere from the early 1970s to the present (Turner, 2014).

An early attempt at building a simple conceptual model of the technosphere itself, particularly its internal structure and dynamics, arose from an analysis of the dynamics of the post-1950 Great Acceleration (Figure 1; Hibbard et al., 2006; Steffen et al., 2007, 2015a). Although developed before the concept of the technosphere was published, this simple conceptual model has several features that are consistent with the technosphere idea and thus may provide a starting point for including it in simple World–Earth System models that represent the coevolutionary dynamics of social-technological macrostructures ('World') and biogeophysical processes ('Earth'). First, the core of the model is a production/consumption loop, driven by energy, which can be linked to a biophysical Earth System model via resource use and waste output. Second, the critical role of science, technology and knowledge (which can include cultural norms and values) in driving the production/consumption loop is explicitly included. Third, the role of human agency via institutions and political economy is included at a scale consistent with the technosphere concept. Although a very simple conceptualisation, this model describes '… a human-created system … that operates beyond our control and that imposes its own requirements on human behaviour' (Haff, 2014a).

Haff's technosphere concept raises important questions about the adequate representation of social and technological mechanisms and constraints in Earth System models. It presents (at least) three basic challenges for current approaches to Earth System modelling:

(1) the technosphere's internal complex dynamics – feedbacks, networks, emergent properties (Verburg et al., 2016) – must be simulated at the global level;
(2) it must be interactively coupled with the rest of the Earth System at the appropriate scale, and its most basic metabolic interactions – the commandeering of energy and resources and

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

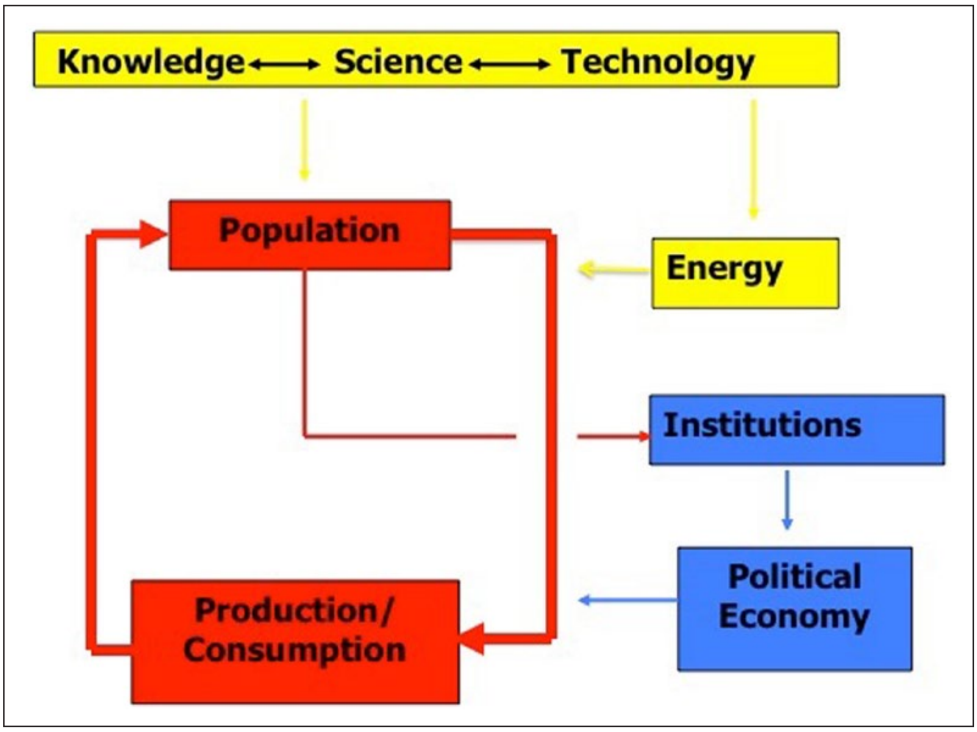


Figure 1. An early conceptual model of the technosphere based on an analysis of the dynamics of the Great Acceleration.
Source: adapted from Hibbard et al. (2006).

the expulsion of waste materials (pollutants) back into the rest of the Earth System – must be simulated; and

(3) the model must account for human (macro-) agency at the appropriate organizational and spatial scale, implying for instance that individual humans cannot influence the technosphere at the scale that matters for Earth System dynamics.

In the following, we discuss adaptive coevolutionary modelling approaches, which might help to tackle these challenges.

The technosphere and emergence of complex coevolutionary dynamics

The technosphere can be thought of as an emergent, coevolved phenomenon of human societies. Issues of scale interaction between the technosphere and human societies should therefore be considered from a coevolutionary perspective. Understanding the emergence of the processes and pathways involved can shed light on the nature of today's interactions between the technosphere and its social sphere of origin. This is particularly important when transitioning from diagnostics of historical developments to projections of possible future trajectories.

From the early palaeolithic, human societies have been characterized by an interwoven complex of technological practices and non-trivial social structures (Camps and Chauhan, 2009). Technologies shaped social structures, while social structures governed the use of technologies (Boserup, 1965). The post-glacial transition from the mesolithic to the neolithic is best understood as a transition in a socio-technological complex (Weisdorf, 2005). Early civilizations were enabled

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

by the technologies they produced and in turn structured by the demands of maintaining these technologies in villages, cities and subsequently across empires (Tainter, 1990). Today, a techno-industrial complex is producing wide-ranging social consequences from the structure of the cities we live in to the channels of communication we use to the daily journeys we undertake. In turn, social dynamics and the resulting systems of preference are continuously influencing the directions and forms technological systems take.

In both cases, the technological and the social, history has seen an emergence of interrelated macro-scale structures. The Great Acceleration of the post-war era should be seen not only as a marked acceleration of the environmental impacts of industrialization, technological innovation, increased global connectivity, availability of energy and the break-through of globalized neoliberal market principles against imperial divisions of territories and practices (Costanza et al., 2007b). It should be equally seen as the more substantial emergence of increasingly large-scale and complex global technological and social structures, namely the technosphere and the human sociosphere (where the word 'sphere' denotes planetary-scale effects). The key question regarding the position of the technosphere in this coevolutionary emergence, today leading to an impending environmental overexploitation of the Earth System with potentially undesirable or even catastrophic outcomes for human societies, is that of collective agency of human societies over the technosphere, as outlined by Haff (2014a).

From the viewpoint of coevolutionary emergence, at issue is both the relationship between agency of individual humans vis-a-vis their social macrostructures such as international institutions, industrial complexes and bureaucratic states, and the collective macro-agency of these macro-scale social entities vis-a-vis the technological macro-infrastructures they collectively have produced and set on their trajectories (Figure 2). Again from a coevolutionary viewpoint, social macrostructures are the product of evolved networks of social interactions. Equally, the technosphere can be conceived of as a network of evolved technological interdependencies, resource and information flows, actions by individuals induced by the technological systems, and interactions with social macrostructures. Haff's particular question on the technosphere concerns the physical and chemical laws governing technological macrosystems. However, since macrosocial and macro-technological complexes have coevolved, there is also a large number of interconnections between the social and technological realms that govern their joint trajectories. The dependencies do not run largely in one direction, from the technological to the social, as Haff implies. Rather, the open question encountered is that concerning joint coevolution, that is, the nature of the coupled interplay between social and technological dynamics.

From the viewpoint of the individual, the challenge is twofold: understanding the relationship of individual agency vis-a-vis macrosocial structures, i.e. the role of the individual as part of an increasingly interconnected mega-society and its institutions, and the interrelationship of these mega-societies with their technocomplexes (Figure 2). In all of this, one should keep in mind that technological realities are heterogeneous across the globe, that historical evolution is spatially asynchronous and shaped by regional preconditions, cultures and preferences. Nonetheless, the present-day dominance of industrialization following a Western development model is striking and seems to be, at least in the present, an attractor of the socio-technological complex once it has emerged.

Modelling the technosphere as adaptive social–technological–ecological networks

To study the technosphere therefore requires three ingredients: consideration of coevolution and emergence, consideration of social, technological and environmental–ecological networks and their coupled macro-dynamics, and considerations of complexity in these dynamics. Only when

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

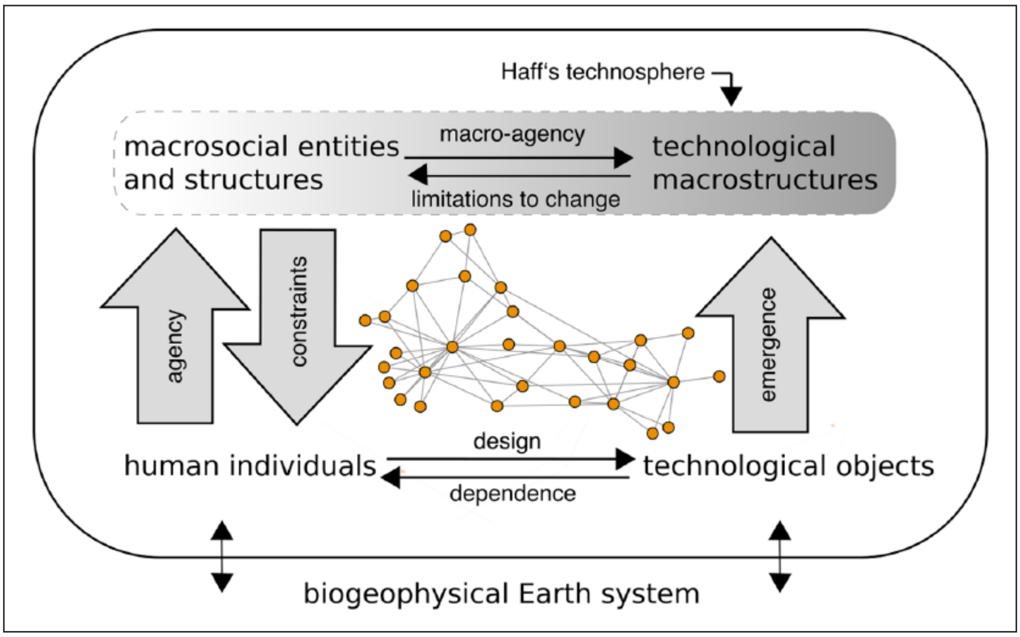


Figure 2. The technosphere reconceptualised as an emergent phenomenon in adaptive social–technological networks in the World–Earth System. The figure illustrates the distinction between individual human agency (micro-level) with respect to influencing macrosocial entities and structures (e.g. nation states, bureaucracies and other social institutions) and their collective macro-agency with respect to technological macrostructures (e.g. the internet, global energy system, industrialized food production). In contrast, Haff's technosphere concept capitalizes on individual human agency mainly with respect to technological macrostructures, but also social macrostructures.

this tangle of coevolutionary effects is somewhat understood would the tools be at hand to ask once more questions about the extent and particular role of human agency in governing the technosphere. To assume that a decoupling of scales occurred between the social and technological macro-levels and the level of individual agency is to downplay the collective effects of a multitude of networks that link the scales. These networks, transferring agency, if indirectly, produce feedbacks between the scales, the overall dynamics of which are hard to predict without the aid of systematic, methodologically sound modelling of complex networks.

We suggest that computer simulation models of the technosphere in an Earth System context as an intertwined social–technological–ecological system should be formulated as adaptive network models (Figure 2; Gross and Blasius, 2008; Gross and Sayama, 2009). These models contain at their core an explicit representation of the coevolutionary dynamics of the states of social, technological and ecological entities (nodes) and their connectivities and interdependencies (links). Within the framework of adaptive coevolutionary networks, social processes such as opinion, preference and coalition formation (Auer et al., 2015; Holme and Newman, 2006; Wiedermann et al., 2015; Schleussner et al., 2016) can be integrated with the metabolic network dynamics of technological infrastructures (Bettencourt et al., 2007; Jarvis et al., 2015) and technological change and innovation, none of which are represented in state-of-the-art Earth System or mainstream integrated assessment models. This perspective is in line with, and should integrate, efforts to apply complex systems approaches and agent-based modelling techniques to the study of the economy (Farmer and Foley, 2009; Farmer et al., 2015) as a key constituent of the technosphere. In such an adaptive network modelling system, human agency would be reflected through decision rules and strategies implemented at different levels of social hierarchy and coarse-graining.

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

The effectiveness of this agency would then be revealed by the degree of their manifestation in the structures and dynamics emerging on macroscopic scales (Figure 2).

Such a modelling effort would need an enormous amount of data and the theoretical knowledge to make use of it, regarding for example the drivers of technological development, government and business decision making on resource use and emissions, and preference formation of consumers (Helbing et al., 2012; Verburg et al., 2016). A big challenge will be to integrate social science research, that operates in case-specific contexts, with the generalizing framework of Earth System models. In view of computational limitations, such models will only work by making significantly simplifying assumptions and generalizations about the complex dynamics of the Earth System, the social metabolisms that operate within it and the environmental and social feedbacks between them. Therefore, we want to stress that the modelling of social–technological systems and, hence, the technosphere, should not aim primarily at prediction of single future development pathways, but at increasing the understanding of their macroscopic properties and emergent dynamics. Such properties of interest include (1) the coarse-grained topology of World–Earth System state space regions of qualitatively different degrees of desirability and safety (including safe and just operating spaces) and resulting management dilemmas (Heitzig et al., 2016); (2) critical control points for the technosphere where human agency can trigger transformative change, e.g. in the energy system; and (3) interactions between social–technological and climatic tipping processes (Schellnhuber, 2009). In this context, it will be relevant to deal with the fact that the self-referentiality of the modeller herself and the infrastructures supporting modelling are parts of the system (the technosphere/Earth System) that she is trying to model. This analytical complication is related to the progression from first order to second order geocybernetics in Earth System analysis as discussed by Schellnhuber (1998).

Conclusions

Reflecting on Haff's technosphere from the point of view of Earth System analysis, we argue on the one hand that in discourses on sustainable development and global change it is highly relevant to take into account explicitly the constraints imposed on human actions by the technosphere (e.g. intrinsic inertia of technological systems), as well as unanticipated risks resulting from feedback dynamics. In addition to environmental risks related to the transgression of planetary boundaries, examples for unpredictable human extinction-level hazards (and related environmental impacts) associated with technological advances including biotechnologies and the emergence of general artificial intelligence (related to the concept of the *singularit*y; Bostrom, 2014) are increasingly coming into the focus of scientific scrutiny as reflected, e.g. by the recent formation of the University of Cambridge Centre for Study for Existential Risks (http://cser.org/). On the other hand, emergent dynamics of the technosphere do not necessarily imply extensive loss of human (macro-) agency as arguably exemplified by the German Energiewende, planned decarbonisation policies in the wake of the Paris 2015 climate agreement, and the social movement on divestment from fossil fuels (Schellnhuber et al., 2016). Consequently, the technosphere should be studied as a coevolutionary planetary phenomenon that can be understood by means of complex systems theory. Computer simulation models as the prominent tools of Earth System analysis play a major role in this endeavour. Therefore, the dynamics of the technosphere and networked feedback processes with the human socio-cultural sphere and the biogeophysical environment need to be captured in next generation models, World–Earth models, to paint a comprehensive panoroma of global sustainability. By allowing a focus on highly relevant emergent critical phenomena such as social–technological tipping elements and their interactions with climatic and biospheric tipping processes, such analytical tools can provide a novel and much needed systemic perspective on the safe and just operating space for humanity and can characterize transformative pathways that lead towards it.

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

Acknowledgements

This paper was developed in the context of project COPAN on Coevolutionary Pathways in the Earth System at the Potsdam Institute for Climate Impact Research (www.pik-potsdam.de/copan). We thank the attendants of the Technosphere-Coevolution working group at the Anthropocene Curriculum in November 2014 organized by the Haus der Kulturen der Welt and the Max Planck Institute for the History of Science, Berlin, Germany as well as the participants of the LOOPS 2015 workshop on ‘From Limits to Growth to Planetary Boundaries: Defining the safe and just space for humanity’ in Southampton, UK for valuable discussions.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: JFD is grateful to the Stordalen Foundation (via the Planetary Boundary Research Network PB.net) and the Earth League’s EarthDoc program for providing financial support. FMH acknowledges funding by the DFG (IRTG 1740/TRP 2011/50151-0).

References

Auer S, Heitzig J, Kornek U et al. (2015) The dynamics of coalition formation on complex networks. *Nature Scientific Reports* 5: 13,386.

Bettencourt LM, Lobo J, Helbing D et al. (2007) Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences of the United States of America* 104(17): 7301–7306.

Boserup E (1965) *The Conditions of Agricultural Growth*. Chicago, IL: Aldine.

Bostrom N (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Camps M and Chauhan PE (eds) (2009) *Sourcebook of Paleolithic Transitions*. New York: Springer.

Costanza R, Graumlich L and Steffen W (eds) (2007b) *Sustainability or Collapse? An Integrated History and Future of People on Earth*. Boston, MA: MIT Press.

Costanza R, Leemans R, Boumans RMJ et al. (2007a) Integrated global models. In: Costanza R, Graumlich L and Steffen W (eds) *Sustainability or Collapse? An Integrated History and Future of People on Earth*. Boston, MA: MIT Press, pp. 417–445.

Elder-Vass D (2010) *The Causal Power of Social Structures: Emergence, Structure and Agency*. New York: Cambridge University Press.

Farmer JD and Foley DK (2009) The economy needs agent-based modelling. *Nature* 460(6): 685–686.

Farmer JD, Hepburn C, Mealy P et al. (2015) A third wave in the economics of climate change. *Environmental and Resource Economics* 62(2): 329–357.

Ferraro F, Pfeffer J and Sutton RI (2005) Economics language and assumptions: How theories can become self-fulfilling. *Academy of Management Review* 30(1): 8–24.

Garrett TJ (2014) Long-run evolution of the global economy: 1. Physical basis. *Earth’s Future* 2(3): 127–151.

Garrett TJ (2015) Long-run evolution of the global economy: 2. Hindcasts of innovation and growth. *Earth System Dynamics* 6(2): 673–688.

Griggs D, Stafford-Smith M, Gaffney O et al. (2013) Policy: Sustainable development goals for people and planet. *Nature* 495(7441): 305–307.

Gross T and Blasius B (2008) Adaptive coevolutionary networks: A review. *Journal of The Royal Society Interface* 5(20): 259–271.

Gross T and Sayama H (2009) *Adaptive Networks*. Berlin, Heidelberg: Springer.

Haff PK (2014a) Humans and technology in the Anthropocene: Six rules. *The Anthropocene Review* 1(2): 126–136.

Haff PK (2014b) Technology as a geological phenomenon: Implications for human well-being. In: Waters CN, Zalasiewicz JA, Williams M, et al. (eds) *A stratigraphical basis for the Anthropocene. Geological Society London, Special Publication* 395: 301–309.

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

Heitzig J, Kittel T, Donges JF et al. (2016) Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system. *Earth System Dynamics* 7(1): 21–50.

Helbing D, Bishop S, Conte R et al. (2012) FuturICT: Participatory computing to understand and manage our complex world in a more sustainable and resilient way. *The European Physical Journal Special Topics* 214(1): 11–39.

Hibbard KA, Crutzen PJ, Lambin EF et al. (2006) Decadal interactions of humans and the environment. In: Costanza R, Graumlich L and Steffen W (eds) *Sustainability or Collapse? An Integrated History and Future of People on Earth*. Boston, MA: MIT Press, pp. 341–375.

Holme P and Newman ME (2006) Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E* 74(5): 056108.

Jarvis AJ, Jarvis SJ and Hewitt CN (2015) Resource acquisition, distribution and end-use efficiencies and the growth of industrial society. *Earth System Dynamics* 6(2): 689–702.

Latour B (2014) Agency at the time of the Anthropocene. *New Literary History* 45(1): 1–18.

Meadows DH, Meadows DL, Randers J et al. (1972) *Limits to Growth*. New York: New American Library.

Moya CJ (1990) *The Philosophy of Action: An Introduction*. Cambridge: Polity Press.

Raworth K (2012) A safe and just space for humanity: Can we live within the doughnut? *Oxfam Policy and Practice: Climate Change and Resilience* 8(1): 1–26.

Ritzer G (2010) *Sociological Theory*. 8th edition. New York: McGraw-Hill.

Rockström J, Steffen W, Noone K et al. (2009) A safe operating space for humanity. *Nature* 461(7263): 472–475.

Schellnhuber HJ (1998) Discourse: Earth System analysis – The scope of the challenge. In: Schellnhuber HJ and Wenzel DV (eds) *Earth System Analysis: Integrating Science for Sustainability*. Berlin/Heidelberg: Springer, pp. 3–195.

Schellnhuber HJ (1999) Earth system analysis and the second Copernican revolution. *Nature* 402: C19–C23.

Schellnhuber HJ (2009) Tipping elements in the Earth System. *Proceedings of the National Academy of Sciences* 106(49): 20,561–20,563.

Schellnhuber HJ, Rahmstorf S and Winkelmann R (2016) Why the right climate target was agreed in Paris. *Nature Climate Change* 6: 649–653.

Schleussner CF, Donges JF, Engemann DA and Levermann A (2016) Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure. *Nature Scientific Reports* 6: 30790.

Schlosser ME (2015) Agency. In: *Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/entries/agency/.

Snijders TA, Van de Bunt GG and Steglich CE (2010) Introduction to stochastic actor-based models for network dynamics. *Social Networks* 32(1): 44–60.

Steffen W, Broadgate W, Deutsch L et al. (2015a) The trajectory of the Anthropocene: The Great Acceleration. *The Anthropocene Review* 2(1): 81–98.

Steffen W, Crutzen PJ and McNeill JR (2007) The Anthropocene: Are humans now overwhelming the great forces of Nature? *Ambio* 36(8): 614–621.

Steffen W, Richardson K, Rockström J et al. (2015b) Planetary boundaries: Guiding human development on a changing planet. *Science* 347: 1259855.

Tainter J (1990) *The Collapse of Complex Societies*. Cambridge: Cambridge University Press.

Turner G (2014) *Is global collapse imminent? An updated comparison of the* Limits to Growth *with historical data*. MSSI Research Paper No. 4, Melbourne Sustainable Society Institute, The University of Melbourne.

Van Vuuren DP, Bayer LB, Chuwah C et al. (2012). A comprehensive view on climate change: Coupling of earth system and integrated assessment models. *Environmental Research Letters* 7(2): 024012.

Verburg PH, Dearing J, Dyke J et al. (2016) Methods and approaches to modelling the Anthropocene. *Global Environmental Change* 39: 328–340.

Weisdorf J (2005) From foraging to farming: Explaining the Neolithic revolution. *Journal of Economic Surveys* 19(4): 561–586.

Wiedermann M, Donges JF, Heitzig J et al. (2015) Macroscopic description of complex adaptive networks coevolving with dynamic node states. *Physical Review E* 91(5): 1–11.

Reproduced from: J. F. Donges et al., The technosphere in Earth System analysis: A coevolutionary perspective, Anthr. Rev. Rev., vol. 4, no. 1, pp. 23–33, 2017, doi: 10.1177/2053019616676608, with the permission of SAGE Publications.

Social tipping dynamics for stabilizing Earth's climate by 2050

Ilona M. Otto[a,1,2], Jonathan F. Donges[a,b,1,2], Roger Cremades[c], Avit Bhowmik[b,d], Richard J. Hewitt[e,f], Wolfgang Lucht[a,g,h], Johan Rockström[a,b], Franziska Allerberger[a,i], Mark McCaffrey[j], Sylvanus S. P. Doe[k], Alex Lenferna[l], Nerea Morán[m,n], Detlef P. van Vuuren[o,p], and Hans Joachim Schellnhuber[a,q,2]

[a]Earth System Analysis, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, 14473 Potsdam, Germany; [b]Stockholm Resilience Centre, Stockholm University, 11419 Stockholm, Sweden; [c]Climate Service Center Germany (GERICS), 20095 Hamburg, Germany; [d]Risk and Environmental Studies, Karlstad University, SE 651 88 Karlstad, Sweden; [e]Information and Computational Sciences Group, James Hutton Institute, Craigiebuckler, Aberdeen AB15 8QH, Scotland, United Kingdom; [f]Observatorio para una Cultura del Territorio, 28012 Madrid, Spain; [g]Department of Geography, Humboldt University, 10099 Berlin, Germany; [h]Integrative Research Institute on Transformations of Human–Environment Systems, Humboldt University, 10099 Berlin, Germany; [i]Department of Geography, University of Innsbruck, 6020 Innsbruck, Austria; [j]UN Climate Change community for Education, Communication and Outreach Stakeholders (ECOS), 3046 Kisbágyon, Hungary; [k]GeoSustainability Consulting, Adabraka-Accra, Ghana; [l]Department of Philosophy, University of Washington, Seattle, WA 98195-3350; [m]Germinando Sociedad Cooperativa Madrid, 28012 Madrid, Spain; [n]Foro de Transiciones, 28011 Madrid, Spain; [o]Climate, Air and Energy, PBL Netherlands Environmental Agency, 2594 AV Den Haag, The Netherlands; [p]Copernicus Institute, Utrecht University, 3584 CB Utrecht, The Netherlands; and [q]Department of Earth System Science, School of Science, Tsinghua University, Haidian District, Beijing 100084, People's Republic of China

Contributed by Hans Joachim Schellnhuber, November 15, 2019 (sent for review January 22, 2019; reviewed by J. David Tabara and Jessika E. Trancik)

Safely achieving the goals of the Paris Climate Agreement requires a worldwide transformation to carbon-neutral societies within the next 30 y. Accelerated technological progress and policy implementations are required to deliver emissions reductions at rates sufficiently fast to avoid crossing dangerous tipping points in the Earth's climate system. Here, we discuss and evaluate the potential of social tipping interventions (STIs) that can activate contagious processes of rapidly spreading technologies, behaviors, social norms, and structural reorganization within their functional domains that we refer to as social tipping elements (STEs). STEs are subdomains of the planetary socioeconomic system where the required disruptive change may take place and lead to a sufficiently fast reduction in anthropogenic greenhouse gas emissions. The results are based on online expert elicitation, a subsequent expert workshop, and a literature review. The STIs that could trigger the tipping of STE subsystems include 1) removing fossil-fuel subsidies and incentivizing decentralized energy generation (STE1, energy production and storage systems), 2) building carbon-neutral cities (STE2, human settlements), 3) divesting from assets linked to fossil fuels (STE3, financial markets), 4) revealing the moral implications of fossil fuels (STE4, norms and value systems), 5) strengthening climate education and engagement (STE5, education system), and 6) disclosing information on greenhouse gas emissions (STE6, information feedbacks). Our research reveals important areas of focus for larger-scale empirical and modeling efforts to better understand the potentials of harnessing social tipping dynamics for climate change mitigation.

climate change | Paris Agreement | decarbonization | social tipping elements | social tipping interventions

Preventing dangerous climate change and its devastating consequences is a defining task for humanity (1, 2). It is also an indispensable prerequisite for achieving sustainable development (3, 4). Limiting global warming to 1.5 °C as stipulated in the Paris Climate Agreement (5) scientifically implies a complete net decarbonization of the world's energy and transport systems, industrial production, and land use by the middle of this century. In their "roadmap for rapid decarbonization," Rockström et al. (6) highlight that rapid increase of the share of zero-carbon energy within the global energy system would be needed to achieve this objective, likely alongside a considerable strengthening of terrestrial carbon sinks. In one scenario, the zero-carbon share of the energy system doubles every 5 to 7 y for the next several decades (6). Carbon emissions that are currently still on the rise at rates of 0 to 2% per year, despite decades-long efforts in international climate negotiations, would thereby need to pivot to a rapid decline of ultimately 7% per year and more. These emission reduction rates would surpass by far even those experienced only during periods of massive socioeconomic crisis in the 20th century, such as World War II and the collapse of communism (Fig. 1).

Here, the historically decisive question is whether and how such rapid rates of deployment can be collectively achieved. Current deployment rates of low-carbon energy sources are compatible with the required shift but when scaled up are expected to encounter considerable resistance due to the rigidities inherent in political and economic decision making (7, 8), as well as new technological demands (9, 10). Although an increasing number of countries have already introduced or are committed to introducing carbon pricing, the initiatives covered by carbon pricing included only 15% of global greenhouse gas emissions in 2017 (11) and have so far driven only marginal emission reductions (12). It is increasingly recognized that business-as-usual technological progress and carbon

Significance

Achieving a rapid global decarbonization to stabilize the climate critically depends on activating contagious and fast-spreading processes of social and technological change within the next few years. Drawing on expert elicitation, an expert workshop, and a review of literature, which provides a comprehensive analysis on this topic, we propose concrete interventions to induce positive social tipping dynamics and a rapid global transformation to carbon-neutral societies. These social tipping interventions comprise removing fossil-fuel subsidies and incentivizing decentralized energy generation, building carbon-neutral cities, divesting from assets linked to fossil fuels, revealing the moral implications of fossil fuels, strengthening climate education and engagement, and disclosing greenhouse gas emissions information.

Author contributions: I.M.O., J.F.D., R.C., A.B., W.L., J.R., and H.J.S. designed research; I.M.O., J.F.D., R.C., A.B., R.J.H., F.A., M.M., S.S.P.D., A.L., N.M., and D.P.v.V. performed research; I.M.O., J.F.D., W.L., and J.R. contributed new reagents/analytic tools; I.M.O., J.F.D., R.C., A.B., and F.A. analyzed data; and I.M.O., J.F.D., R.C., A.B., R.J.H., W.L., J.R., F.A., M.M., S.S.P.D., A.L., N.M., D.P.v.V., and H.J.S. wrote the paper.

Reviewers: J.D.T., Autonomous University of Barcelona; and J.E.T., Massachusetts Institute of Technology.

The authors declare no competing interest.

[1]I.M.O. and J.F.D. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: ilona.otto@pik-potsdam.de, donges@pik-potsdam.de, or john@pik-potsdam.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1900577117/-/DCSupplemental.

First published January 21, 2020.

Reproduced from: I. M. Otto et al., Social tipping dynamics for stabilizing Earth's climate by 2050, Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 5, pp. 2354–2365, 2020, doi: 10.1073/pnas.1900577117. Published under Creative Commons Attribution License 4.0 (CC BY).

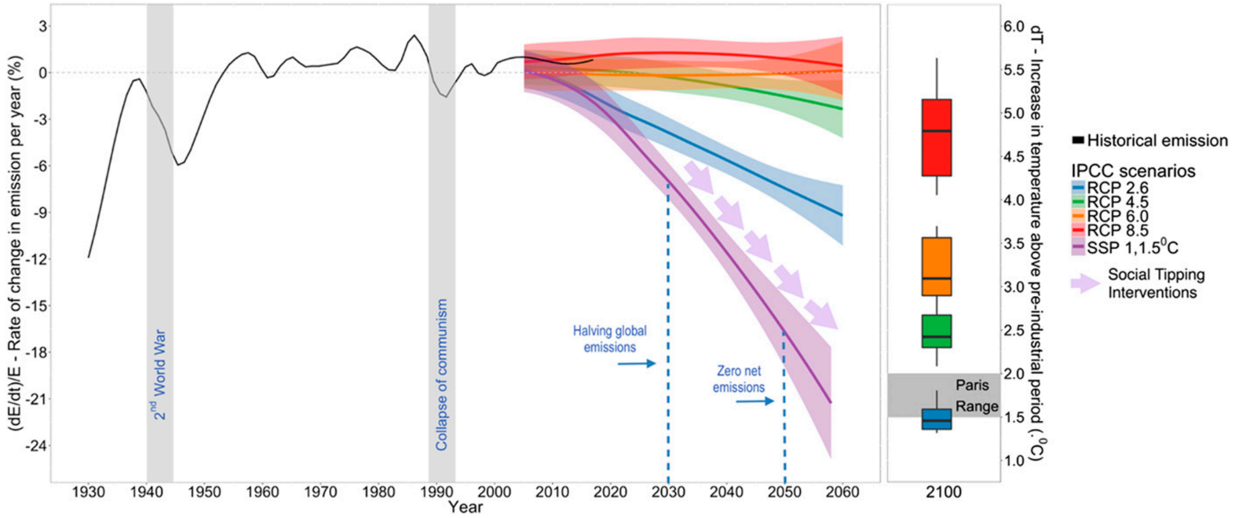


Fig. 1. The rate of change in annual greenhouse gas emissions required for net decarbonization. Social tipping dynamics in the context of the representative concentration pathways (RCPs) of the Intergovernmental Panel on Climate Change (IPCC) and the Paris Agreement. *Left* and *Right* exhibit the rate of change in CO_2 emission per year between 1930 and 2060, and the increase in global mean temperature by 2100 relative to the preindustrial period, respectively, under the four RCPs. The transition to a new net decarbonized state requires shifting from an incremental rise in emissions of 0 to 2% per year to nonlinear decline at the rate of 7% per year and more (6). The figure was created using the RCP emission projections (153) and Coupled Model Intercomparison Project 5 (CMIP5) temperature projections (154).

pricing alone are not likely to lead to rapid and deep reductions in greenhouse gas emissions (13).

At the same time, there is evidence from various scientific fields demonstrating that rapid rates of change can be observed under certain critical conditions in natural (14–16), socioeconomic (17–20) and social-ecological systems (SESs) (21, 22). Increasing attention is being given to the concept of tipping dynamics as a nonlinear mechanism behind such disruptive system changes. Based on a review on social-ecological tipping points research, Milkoreit et al. (23) propose a common definition of social tipping points (STPs) as points "within an SES at which a small quantitative change inevitably triggers a non-linear change in the social component of the SES, driven by self-reinforcing positive-feedback mechanisms, that inevitably and often irreversibly lead to a qualitatively different state of the social system." There are historical examples of dynamic social spreading effects leading to a large self-amplification of small interventions: For example, the writings of one man, Martin Luther, injected through newly available printing technology into a public ready for such change, triggered the worldwide establishment of Protestant churches (24). An example in the field of climate policy is the introduction of tariffs, subsidies, and mandates to incentivize the growth of renewable energy production. This has led to a substantial system response in the form of mutually reinforcing market growth and exponential technology cost improvement (25, 26).

In this paper, we examine a number of potential "social tipping elements" (STEs) for decarbonization (27, 28) that represent specific subdomains of the planetary social-economic system. Tipping of these subsystems could be triggered by "social tipping interventions" (STIs) that could contribute to rapid transition of the world system into a state of net zero anthropogenic greenhouse gas emissions. The results reported in this study are based on an online expert survey, an expert workshop, and an extensive literature review (*SI Appendix*).

Our results complement the existing shared socioeconomic pathways (SSPs) that are used alongside the representative concentration pathways (RCPs) to analyze the feedbacks between climate change and socioeconomic factors, such as world population growth, economic development, and technological progress (29). Our results could be useful for exploring possible transformative pathways leading to scenarios that reach net zero emissions by 2050 (30).

Defining STEs and STIs Relevant for Decarbonization Transformation

Various types of tipping processes can be differentiated in the literature. Many authors refer to critical thresholds (16, 28), a notion closely related to the metaphor of a "butterfly effect" (31, 32). Other processes related to tipping dynamics include metamorphosis, where a rapid loss of structures of one sort occurs simultaneously with the development of new structures (33), as well as cascades driven by positive feedbacks in processes occurring simultaneously at smaller scales (34).

The social tipping dynamics of interest for this study are typically manifested as spreading processes in complex social networks (35, 36) of behaviors, opinions, knowledge, technologies, and social norms (37, 38), including spreading processes of structural change and reorganization (34). These spreading processes resemble contagious dynamics observed in epidemiology that spread through social networks (39). Once triggered, such processes can be irreversible and difficult to stop. Similar contagious dynamics have been observed in human behavior (35, 36), for example in assaultive violence (39), participation in social movements (40), or health-related behaviors and traits (36), such as smoking or obesity (41, 42).

We understand STEs as functional subsystems of the planetary-scale World–Earth system (43) consisting of interacting biophysical subsystems of the Earth, and the social, cultural, economic, and technological subsystems of the world of human societies (43, 44). Potential STEs share one defining characteristic: A small change or intervention in the subsystem can lead to large changes at the macroscopic level (23) and drive the World–Earth system into a new basin of attraction, making the transition difficult to reverse (20). Exact quantifications of the relationship between big and small are, however, rare, as are empirical examples (Table 1). For the combination of big interventions and big effects, there are currently no convincing examples; however, the potential use

Table 1. Illustrative examples of intervention-and-effect relationships in the context of climate change mitigation

Intervention types	Small effect	Big effect
Small intervention	An incremental change, e.g., a town mitigation plan (157)	A tipping effect, e.g., feed-in tariffs in the German "Energiewende" (158)
Big intervention	Inefficient interventions, e.g., the implementation of the European Carbon Emission Trading Scheme leading to a marginal reduction of greenhouse gas emissions due to leakage effects (159)	An elephant effect, e.g., reducing the Earth's carbon burden by means of solar radiation management geoengineering (160)

solar radiation management geoengineering in the future would fall into this category. Finally, some changes in the World–Earth system might be driven by nonhuman and unintentional forces (e.g., a sufficiently large meteorite hitting the Earth or a disease outbreak), while others might be driven by conscious interventions of human agency (45).

Tipping processes might be analyzed as a function of change in a suitably selected forcing variable or control parameter (15, 27). The pertinent World–Earth system features such as the anthropogenic carbon emissions are commonly the product of complex interactions of multiple drivers. These factor can, however, in some cases be combined into a single dominant control parameter.

In this study, we identify a subsystem of the World–Earth system as a STE relevant for decarbonization transformation if it fulfils the following criteria:

C1. A set of parameters or drivers controlling its state can be described by a combined control parameter that after crossing a critical threshold (the STP) by a small amount influences a crucial system feature of relevance (here the rate of anthropogenic greenhouse gas emissions) leading to a qualitative change in the system after a reference time has passed allowing for the emergence of the effect (15).

C2. It is possible to differentiate particular human interventions leading to the small change in the control parameter that has a big effect on the crucial system feature, which will be referred to as the STI (Fig. 2).

Established social systems, including their infrastructures, while they may partly be open to change, tend also to possess self-stabilizing mechanisms that oppose change, be it through infrastructural inertia due to investment cycles or cultural or political inertia due to deeply held traditions or power structures all representing aspects of social complexities (Fig. 2 and refs. 46 and 47). For this reason, a cumulation of effects due to social contagion, repetitive nudging, or direct intervention can lead to social tipping dynamics (48). Starting points for such cumulations of effects are here called STIs. Naturally, their existence, nature, and point of departure are a function of the cumulated history of the respective social system and, in that sense, STIs and social tipping dynamics are path dependent.

Following Rockström et al. (6), in order to achieve the Paris Climate Agreement's goals and to avoid higher levels of global warming at the end of this century that would imply crossing dangerous tipping points in the Earth's climate system (27), global anthropogenic carbon emissions would need to be halved every decade, achieving a peak in 2020 and then steadily decreasing to reach net zero emissions by 2050. Achieving net zero global emissions around 2050 is necessary for there to be a significant probability of limiting global warming to 1.5 °C by the end of the century (1). To ensure that the social tipping dynamics identified in this study are compatible with these constraints, we impose the following filtering criteria:

F1. The time needed to trigger the tipping should not exceed ~15 y, and the time needed to observe a qualitative change at the whole system level should not exceed ~30 y (Fig. 1).

F2. Since abrupt social changes have historically often been associated with social unrest, war, or even collapse (49), human intervention and its foreseeable effects should here be explicitly compatible with the Sustainable Development Goals (50), in the sense of positive social tipping dynamics (34).

Finally, due to the networked and multilevel character of the social system (51), we also ask about the feedback mechanisms connecting and potentially mutually reinforcing the identified candidates for STEs and STIs.

Results

Candidates for STEs from Expert Elicitation. Both natural and social systems are characterized by a high level of complexity and are linked by coevolutionary dynamics (52). Isolating the elements of such systems is difficult. Although we provided our respondents with a written definition of a STE, most of the online survey participants referred to what we define as STIs. On the basis of the responses, 12 groups of candidates for STEs could be identified, each referring to a distinctive control parameter (Table 2). The critical threshold of the control parameter needed to be crossed in order to trigger the tipping process was in most of groups not quantified by the experts but described qualitatively. The STP was often referred to as the point when a certain belief, behavior, or technology, spreads from a minor tendency to a major practice. Documented instances of technology and business solutions show that a 17 to 20% market or population share can be sufficient to cross the tipping point and scale up to become the dominant pattern (53). Some authors, however, argue that it must be the "right" share of population, including well-connected influential people, trendsetters, and other types of social leaders with a high degree of agency (38, 54). In other cases, the experts referred to the STP that would be achieved if the price of fossil-fuel–free products and services falls below that of those products and services based on fossil fuels. Table 2 presents an overview of expert elicitation results.

Critical Interventions for Inducing Social Tipping Dynamics. Building upon the results of our expert elicitation, we differentiated six key candidates for STEs and associated STIs for which we were able to find empirical material showing that they fulfill the conditions specified in our definition (as listed in Table 3). These do not necessarily comprise a comprehensive list of "silver bullet" solutions; rather, this is an initial selection that can help in developing more refined socioeconomic rapid transformation pathways and narratives customized at appropriate scales. Below, we present a review of literature on each of the STEs and STIs nominated by the experts. We search for evidence supporting the potential of the interventions to trigger tipping-like changes in their domains leading to a qualitative change at the World–Earth system level; we ask whether critical thresholds in

Reproduced from: I. M. Otto et al., Social tipping dynamics for stabilizing Earth's climate by 2050, Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 5, pp. 2354–2365, 2020, doi: 10.1073/pnas.1900577117. Published under Creative Commons Attribution License 4.0 (CC BY).

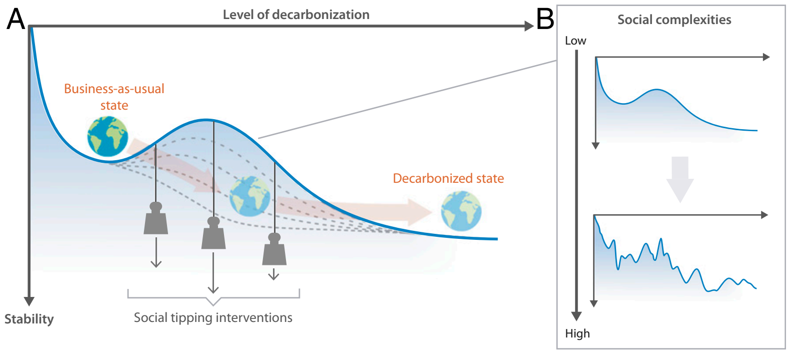


Fig. 2. The concept of decarbonization transformation as social tipping dynamics. As illustrated in *A* by an abstract stability landscape (155), the world's socioeconomic system today is trapped in a valley where it still depends heavily on burning fossil fuels, leading to high rates of greenhouse gas (GHG) emissions. STIs have the potential to erode the barrier through triggering social tipping dynamics in different sectors (Fig. 3) and thus paving the way for rapid transformative change. Uncertainties and complexities inherent in the many dimensions of human societies beyond their level of decarbonization (46) can be envisioned as forming a rougher stability landscape featuring multiple attracting states and a larger number of barriers that need to be eroded or overcome (*B*). This inherent "social noise" may complicate transformative change but could also accelerate it by means of dynamical phenomena such as stochastic resonance (156).

the control parameters can be determined; and finally we begin to examine the interactions and feedbacks among the identified tipping elements.

STIs in the energy production system. The technological development in the energy production system is a dominant element of the decarbonization discussions in international institutions (55, 56) and business partnerships (57). The results of our expert elicitation confirm that technology development is likely to play a key role, however, not in the sense of yet-to-be invented technological solutions, but rather in the adaptation of existing carbon-free technology primarily in the power sector and by facilitating a smarter utilization of energy. The main control parameter that drives the adaptation of fossil-fuel–free energy technology is associated with the financial returns of its adoption (58). Our expert group believed that the critical condition needed to trigger the tipping process is the moment when fossil-fuel–free energy production yields higher financial returns than the energy production based on fossil fuels. The empirical data show that this critical threshold is about to be reached; the prices of renewables have dropped sharply in the last few years, and they have already become the cheapest source of energy in many world regions. The average cost of onshore wind dropped by 18%, and offshore wind fell by 28% (59). The costs of photovoltaic modules fell by about 20% with every doubling of cumulative capacity since the 1970s (60) and the key role in reducing the cost of photovoltaics was played by policies that stimulate market growth (26). Optimization modeling shows that renewable energy supplies can potentially supply 100% of human power demand (61), and in theory, rapid transformation to low energy demand is possible (30) and will be cost-effective in the long run (62). However, there are large costs associated with adapting existing infrastructure and supply and demand support services to meet the demands of nondispatchable, volatile renewable sources like wind and solar in electricity generation. The question is whether the cost of transforming the energy infrastructure is worthwhile compared to the cost of inaction. The prioritization of societal preferences in the competition for scarce budgetary resources is influenced by the dominant social values (63).

Our expert group believed that redirecting national subsidy programs to renewables and low-carbon energy sources or removing the subsidies for fossil-fuel technologies are the tipping interventions that are needed for the take-off and diffusion of fossil-fuel–free energy systems. The key actors who have the agency to implement these interventions include national governments and energy ministries, and the response of large energy companies is important. One-third of global industrial greenhouse gas emissions can be linked just to 29 oil and gas companies (64). The International Energy Agency has tracked fossil-fuel subsidies over the last decade and in 2009 estimated that $312bn was spent worldwide in fossil-fuel subsidies, compared to $57bn on renewables in that year (65). By 2015, the gap had narrowed, but the subsidies received by fossil fuels were still more than twice those of renewables (66). Estimates show that a universal phaseout of fossil-fuel subsidies could lower annual carbon emissions by 4.4% (67). Coady et al. (68) argue that eliminating subsidies for fossil fuels would have reduced global carbon emissions in 2013 by 21%.

Furthermore, our expert group believed that the global energy production and storage system can also be radically changed by decentralization of energy production. Since large power stations relying on coal, oil, or gas exploitation are not profitable below a certain threshold of households supplied, decentralized generation systems and transitioning to local power generation might be expected to lead to a virtually complete decarbonization of production systems (69, 70). However, this is also likely to lead to an increase in costs due to the loss of economies of scale (69), and the complexities of integrating variable, distributed power sources (71). This emphasizes the need for decentralized energy generation and demand management to be part of the wider energy systems transformation (72). It has been argued that citizens also have a major role to play as nodes in a smart system capable of facilitating flexible demand management (73). Some authors also warn that meeting current levels of demand (let alone future projected demand) with renewables alone is likely to be extremely difficult (74, 75). Nonetheless, interest in decentralized control of energy systems is growing. Across the Global North, there are a multitude of examples of energy cooperatives and community-driven energy projects (76). Such projects have often found creative ways to overcome limitations imposed by centralized distribution networks, e.g., by using smart technologies to divert excess power for local heating (77), or by bringing municipal supply networks into community ownership (78). They show such initiatives may also spark around the Global South by skipping the "megadevelopment" phase associated with large power stations and massive grid infrastructure expansion. Due to the positive knowledge and technology spillover effects from such decentralized systems, the technology costs are likely to be further reduced with their increased diffusion (79, 80). The time elapsing between the planning phase and actual installation and utilization of decentralized energy generation is reportedly less than 10 y (81).

Reproduced from: I. M. Otto et al., Social tipping dynamics for stabilizing Earth's climate by 2050, Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 5, pp. 2354–2365, 2020, doi: 10.1073/pnas.1900577117. Published under Creative Commons Attribution License 4.0 (CC BY).

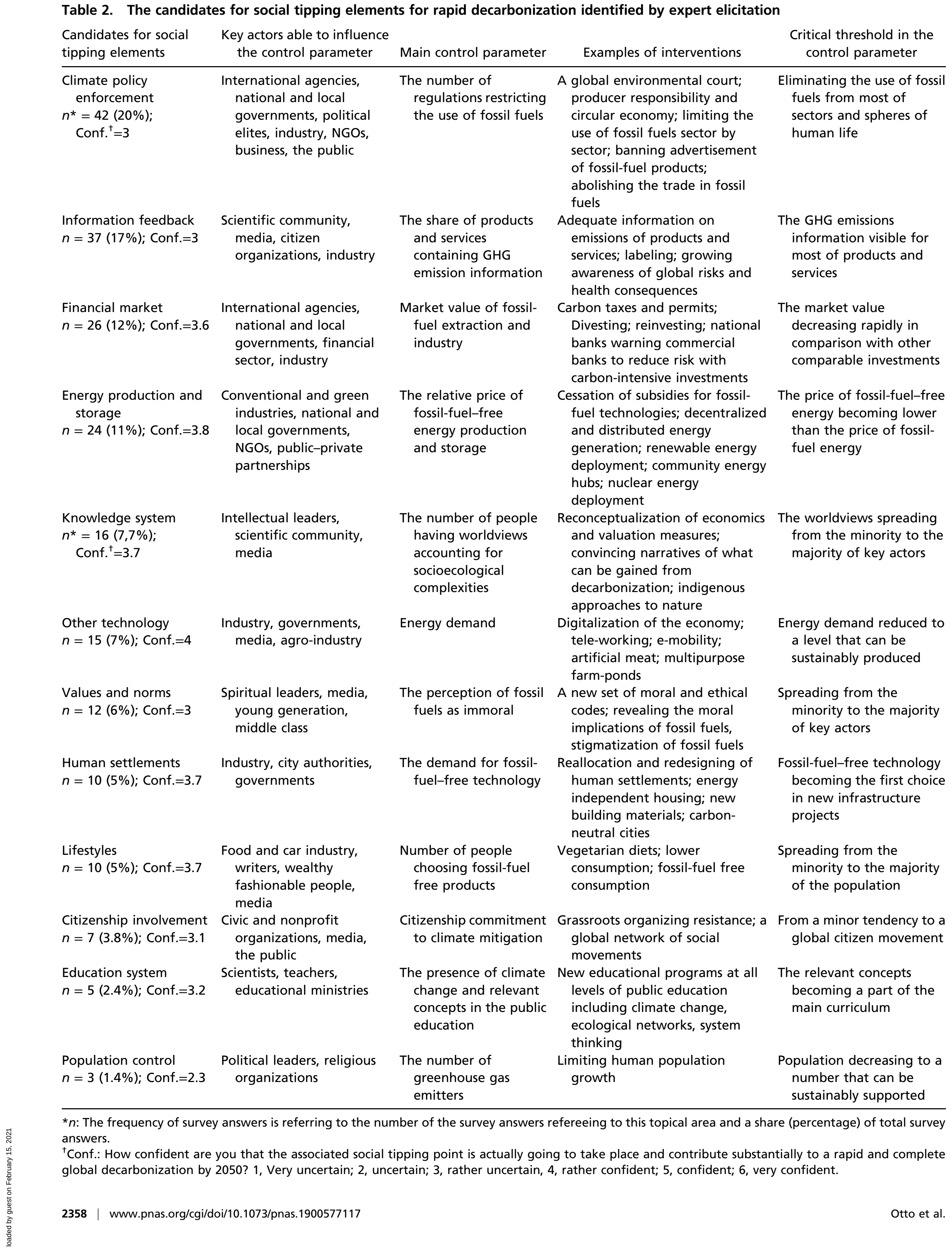


Table 2. The candidates for social tipping elements for rapid decarbonization identified by expert elicitation

Candidates for social tipping elements	Key actors able to influence the control parameter	Main control parameter	Examples of interventions	Critical threshold in the control parameter
Climate policy enforcement *n** = 42 (20%); Conf.[†]=3	International agencies, national and local governments, political elites, industry, NGOs, business, the public	The number of regulations restricting the use of fossil fuels	A global environmental court; producer responsibility and circular economy; limiting the use of fossil fuels sector by sector; banning advertisement of fossil-fuel products; abolishing the trade in fossil fuels	Eliminating the use of fossil fuels from most of sectors and spheres of human life
Information feedback *n* = 37 (17%); Conf.=3	Scientific community, media, citizen organizations, industry	The share of products and services containing GHG emission information	Adequate information on emissions of products and services; labeling; growing awareness of global risks and health consequences	The GHG emissions information visible for most of products and services
Financial market *n* = 26 (12%); Conf.=3.6	International agencies, national and local governments, financial sector, industry	Market value of fossil-fuel extraction and industry	Carbon taxes and permits; Divesting; reinvesting; national banks warning commercial banks to reduce risk with carbon-intensive investments	The market value decreasing rapidly in comparison with other comparable investments
Energy production and storage *n* = 24 (11%); Conf.=3.8	Conventional and green industries, national and local governments, NGOs, public–private partnerships	The relative price of fossil-fuel–free energy production and storage	Cessation of subsidies for fossil-fuel technologies; decentralized and distributed energy generation; renewable energy deployment; community energy hubs; nuclear energy deployment	The price of fossil-fuel–free energy becoming lower than the price of fossil-fuel energy
Knowledge system *n** = 16 (7,7%); Conf.[†]=3.7	Intellectual leaders, scientific community, media	The number of people having worldviews accounting for socioecological complexities	Reconceptualization of economics and valuation measures; convincing narratives of what can be gained from decarbonization; indigenous approaches to nature	The worldviews spreading from the minority to the majority of key actors
Other technology *n* = 15 (7%); Conf.=4	Industry, governments, media, agro-industry	Energy demand	Digitalization of the economy; tele-working; e-mobility; artificial meat; multipurpose farm-ponds	Energy demand reduced to a level that can be sustainably produced
Values and norms *n* = 12 (6%); Conf.=3	Spiritual leaders, media, young generation, middle class	The perception of fossil fuels as immoral	A new set of moral and ethical codes; revealing the moral implications of fossil fuels, stigmatization of fossil fuels	Spreading from the minority to the majority of key actors
Human settlements *n* = 10 (5%); Conf.=3.7	Industry, city authorities, governments	The demand for fossil-fuel–free technology	Reallocation and redesigning of human settlements; energy independent housing; new building materials; carbon-neutral cities	Fossil-fuel–free technology becoming the first choice in new infrastructure projects
Lifestyles *n* = 10 (5%); Conf.=3.7	Food and car industry, writers, wealthy fashionable people, media	Number of people choosing fossil-fuel free products	Vegetarian diets; lower consumption; fossil-fuel free consumption	Spreading from the minority to the majority of the population
Citizenship involvement *n* = 7 (3.8%); Conf.=3.1	Civic and nonprofit organizations, media, the public	Citizenship commitment to climate mitigation	Grassroots organizing resistance; a global network of social movements	From a minor tendency to a global citizen movement
Education system *n* = 5 (2.4%); Conf.=3.2	Scientists, teachers, educational ministries	The presence of climate change and relevant concepts in the public education	New educational programs at all levels of public education including climate change, ecological networks, system thinking	The relevant concepts becoming a part of the main curriculum
Population control *n* = 3 (1.4%); Conf.=2.3	Political leaders, religious organizations	The number of greenhouse gas emitters	Limiting human population growth	Population decreasing to a number that can be sustainably supported

**n*: The frequency of survey answers is referring to the number of the survey answers refereeing to this topical area and a share (percentage) of total survey answers.

[†]Conf.: How confident are you that the associated social tipping point is actually going to take place and contribute substantially to a rapid and complete global decarbonization by 2050? 1, Very uncertain; 2, uncertain; 3, rather uncertain, 4, rather confident; 5, confident; 6, very confident.

Reproduced from: I. M. Otto et al., Social tipping dynamics for stabilizing Earth's climate by 2050, Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 5, pp. 2354–2365, 2020, doi: 10.1073/pnas.1900577117. Published under Creative Commons Attribution License 4.0 (CC BY).

Table 3. Synthesis of the research results on the key candidates for social tipping elements selected by the experts and their associated social tipping interventions

Social tipping element	Social tipping intervention	Control parameter	Key actors	GHG emission reduction potential	Dominant social structure level	Estimated time needed to trigger tipping
STE1: Energy production and storage	STI1.1: Subsidy programs	The relative price of fossil-fuel–free energy	Governments, energy ministries, big energy producers (68)	Up to 21% globally in 1 y (68)	National policy (68)	10 to 20 y (including the policy-formative phase) (161)
	STI1.2: Decentralized energy production		Citizens, communities (73), local governments (162), policy makers (163), energy planners (164)	Up to 100% in power supply (61)	Community/town governance (165)	Less than 10 y (81)
STE2: Human settlements	STI2.2: Carbon-neutral cities	The demand for fossil-fuel–free technology	City administration, citizens, and citizen groups (166)	Reduction by 32% in 14 y (91)	Urban governance (91)	Approximately 10 y (91).
STE3: Financial market	STI3.1: Divestment movement	Profitability of fossil fuel exploitation	Financial investors (96)	26% emissions tied to investments of a large Canadian university (167)	Market exchange, enterprise (98)	Very rapid, could occur within hours (142)
STE4: Norms and values system	STI4.1: Recognition of the immoral character of fossil fuels	The perception of fossil fuels as immoral	Peer groups, environmental organizations, youth, opinion leaders (168–170)	Unprecedented	Informal institutions, enforcement through peer groups (171)	30 to 40 y (172)
STE5: Education system	STI5.1: Climate education and engagement	Climate change and impacts awareness	Teachers, climate educators (117), youth (113)	Up to 30% reduction in 2 y in the emissions of the Italian households included in the study (124)	National policy (173)	10 to 20 y (173)
STE6: Information feedback	STI6.1: Emission information disclosure	The number of products and services disclosing their carbon emissions	The business and producers; governments for setting disclosure guidelines and regulations (174)	Up to 10% reduction of emissions in UK households' grocery consumption in a year (175)	Market, exchange (176); enterprise (177)	A few years (178)

However, existing energy systems and infrastructure are likely to shape the future for decades to come (82).

STIs in human settlements. Direct and indirect emissions from buildings account for almost 20% of all carbon emissions, and we observe an unprecedented scale of global urbanization; each week the global urban population increases by 1.3 million (55). The average life span of buildings is about 50 y (83). Public infrastructure and planning structures can last even longer (50 to 150 y) and play an active role in both climate mitigation and adaptation (84). Modifying building codes for construction and infrastructural projects can actively drive the demand for fossil-fuel–free technologies and are crucial especially for countries in the Global South, where building booms are driving up energy and other resource use (85). An example of a STI in this realm is the creation of large-scale demonstration projects such as carbon-neutral cities. These are important in order to educate the general public and stimulate consumer interest in environmental technologies, accelerating their dissemination and commercialization (85). In addition, local technology clusters create positive spillover effects of lowering the information and transaction costs (86), which can indirectly lead to a reduction in the costs of fossil-fuel–free technologies for energy production and storage. The critical conditions for social tipping in this control parameter would be achieved if the fossil-fuel–free technology became the first choice for new construction and infrastructure projects. There are many new construction materials that not only imply lower emissions but also could actively support carbon sequestration efforts in urban areas. To give an example, constructing a 142-m-high residential building using above 80% laminated timber could lead to sequestrating 21,040 tons CO_2 and avoiding 50,000 tons CO_2 emissions otherwise entailed in using standard construction materials such as steel and concrete, which is equivalent to the amount 33,000 cars emit per year (87). In addition, large-scale public infrastructure investments support the emergence of a shared belief in the emerging new social equilibrium that can help individuals coordinate changes and find new focal points (88). The example of the Transition Town Movement that started in 2006 in the United Kingdom and in 2014 spanned over 41 countries shows how local grassroots initiatives can encourage citizens to take direct action toward lowering energy demand and building local resilience despite lack of policy support at national levels (89). Another example includes the Energy Cities Association, whose primary goal is to accelerate the transition to sustainable energy in urban areas in Europe. The Association was created in 1990 and currently represents more than 1,000 towns and cities in 30 countries (90). The evidence from a case study on communities implementing plans for zero

Reproduced from: I. M. Otto et al., Social tipping dynamics for stabilizing Earth's climate by 2050, Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 5, pp. 2354–2365, 2020, doi: 10.1073/pnas.1900577117. Published under Creative Commons Attribution License 4.0 (CC BY).

emissions shows that these communities were able to reduce their per-capita emissions by 32% in 14 y (91).

STIs in the financial system. The financial crisis in 2008 demonstrated how rapidly changes in the market value of assets in one sector and country can propagate and destabilize the global system of human societies and accelerate changes at the level of individual investment and consumption behavior as well as collective-organizational and policy responses (92). Maintaining global warming below 2 °C implies that 33% of oil, 49% of gas, and 82% of coal resources should not be burned (93). This suggests there might be a risk of a carbon bubble, caused by the financial exposure from stranded assets, which could be driven by policy, technological innovation, or investors' decisions (94). A growing number of analysts believe a financial bubble is emerging that could burst when investors' belief in carbon risk reaches a certain threshold (95). Simulations show that just 9% of investors could tip the system, inducing other investors to follow (96). An example of an intervention that can lead to a rapid decline in the control parameter—the value of fossil-fuel assets—is the divestment movement; as it progresses, it results in the reduction of the value of fossil-fuel assets (97). The movement started with a student campaign in 2011 and is quickly expanding to other countries and types of asset owners. The value of investment funds committed to selling off fossil-fuel assets reached $5.2tn in 2016, doubling in just over a year and permeating enterprises in every sector of society, with examples including universities, faith groups, pension funds, and insurance companies (98). Ritchie and Dowlatabadi (94) present model scenarios showing that a major Canadian university could reduce the greenhouse gas emissions tied to its investments by up to 26% by restructuring its portfolios, moving investments away from greenhouse gas-intensive sectors. Many divestment campaigns have an additional "divest to reinvest" element that advocates using funds invested in fossil-fuel companies to reinvest in socially and environmentally beneficial projects, such as low-carbon and renewable schemes (99), creating the positive-feedback interactions with the STE1. An avalanche effect would be triggered if national banks and insurance companies warned against the global risk associated to stranded assets from fossil-fuel projects. These concerns are growing in Europe, and there are already signs of a tipping point, namely cuts in financial and insurance support for coal projects (100). Norwegian financial authorities might soon be divesting the country's sovereign wealth fund. Around 6% (€30bn) of this fund's wealth is invested in oil and gas companies (101).

STIs in the system of norms and values. The extraction and use of fossil fuels out of line with the Paris Climate Agreement targets is arguably immoral, as it would cause widespread grave and unnecessary harm (97). The impact of greenhouse gas emissions disproportionately affects the most vulnerable social groups, such as women and children (102). It also affects the well-being of future human generations (103) and causes many direct negative health effects (104). Historical cases show that social and moral norms can affect human behavior on a large scale (38). The abolition of the transatlantic slave trade, for example, showed that changes in the ethical perception of slave labor at that time were consciously initiated by a small group of intellectuals (105). Revealing the moral implication of the continued burning of fossil fuels is an example of an intervention that is likely to induce a tipping process through changes in the human normative system, i.e., the system of moral and behavioral norms that influence what is rewarded and desired in the society. Norms can develop through social networks in neighborhoods or workplaces and support certain lifestyles or technology choices (106). A study on the installation of photovoltaic panels by home owners showed social networks and dwelling proximity explained the owners' decision to install photovoltaic panels on their homes (107). The control parameter is represented by the ethical perception of fossil fuels, the environmental externalities they generate, and the broader harm they visit on societies. The critical condition in the control parameter will be achieved if the majority of social and public opinion leaders recognize the ethical implications of fossil fuels and generate pressure in their peer groups to ostracize the use of products involving fossil fuel burning. This could be more widespread in religious communities and be led by spiritual leaders, perhaps following the example of Pope Francis's encyclical *Laudato si'* (108). It could alternatively be manifested as a secular trend originating mainly from young, intellectually and social justice-oriented groups of people who might actively stand against supporters of fossil fuels—these would include extraction and utilization companies, governments supporting the latter, as well as the superrich family clans generating wealth from fossil fuel extraction and utilization in the last 150 y. The wealth of about 11% of the world's billionaires is related to energy production (excluding solar and wind), mining, and other natural resource utilization (109). Recent experimental evidence shows that dominant social conventions or established behavior can be changed by committed minorities of roughly 25% of a group (36). Social norms are the sources of law (110); therefore, recognizing the immoral character of fossil fuels can further lead to regulations restricting the use and extraction of fossil fuels (111).

The time elapsing from the recognition of the activity as a problem and as a matter of a moral choice by international legal scholars, religious groups, and other moral entrepreneurs, to international delegitimization might range from a few decades to a few centuries. The slavery abolition movement started in 1772 in England and led to the abolition of the slave trade in 1807 and in the 1833 to the total abolition of slavery in the British Empire. The historical data show that although the number of slaves traded in the British Empire dropped to zero by 1826, the number of internationally traded slaves started to decrease around the mid-19th century. However, after reaching its peak, the number of slaves traded internationally decreased exponentially within just a few years. In the period 1851 to 1860, 71% fewer slaves disembarked than in the period 1841 to 1850 (https://slavevoyages.org/). A more recent example of outlawing the use of substances responsible for ozone depletion showed that such changes might occur in less than 30 y (112). However, the financial and political power of the fossil fuel industry suggests the need for much more substantial political effort to ensure such a change, than would have been the case for the issue of ozone depletion (99). There is recent anecdotal evidence that protests, such as the #FridaysForFuture climate strikes of school students around the world, the Extinction Rebellion protests in the United Kingdom, and initiatives such as the Green New Deal in the United States, might be indicators of this change in norms and values taking place right now (113).

STIs in the education system. Many examples of research confirm the role of education in social transformations (114) and tackling climate change concerns (115, 116). The control parameter that relates to this intervention is the coverage of climate change issues in school and university teaching programs. While many teachers include some, often thin, coverage of climate change (117), comprehensive approaches at all levels of public education are still rare. Lack of knowledge about the causes, impacts, and solutions to climate change was the most easily identifiable individual barrier to engagement in climate action in the United Kingdom (118). At the same time, studies show that the divergent ways of understanding climate change draw on discourses broader than scientific knowledge; these differences may be blamed for misinterpretation of scientific notions such as uncertainty (119) as well as for the tendency to attribute responsibility for causing and mitigating climate change to others (118). Formal and lifelong education is traditionally considered a slow and evolving process, but there are examples of rapid change that can be generated. Quality education supports and amplifies norms and values and can quickly inspire behavior change among individuals and their cohorts. In addition, massive literacy campaigns, such as the one that took place in Cuba in the 1950s, where in a less than a year illiteracy was reduced from 24 to 3.9% (120), demonstrate the potential for rapid societal transformation. The effects of changes in educational programs can also lead to a social tipping process as soon as the new generation enters the job market and public decision-making bodies. The recent #FridaysForFuture protests demonstrate the upcoming new generation might radically change the political scene. It is estimated that within just half a year the school children movement grew to 1.5 million students in 125 countries. The effects of educational campaigns can be strengthened by a supportive family and community context as well as by media campaigns, advertising bans, higher taxes, use prohibitions, and lawsuits against producers (121). Warner (122) shows that combined educational and mass-media campaigns in the 1970s in the United States led to 4 to 5% annual decrease in cigarette consumption. In the climate change context, Dietz et al. (123) show that interventions that combine mass-media messages, household- and behavior-specific information, and communication through individuals' social networks and communities could lead to reductions of 20% in household direct emissions in less than 10 y, with little or no reduction in household well-being. An educational campaign carried out in five Italian cities for 2 y, involving teachers, pupils, and citizens, resulted in an emission reduction in a range of 7 to 30% in the 247 families included in the research (124). That said, education to bolster understanding of the causes and effects of climate change, however important, will not be sufficient to transform society alone. Sustainability cannot be imposed, it has to be learned, so that is endogenously realized and enacted deliberately by the actors who constitute the SES (46). Engagement and the fostering of sustainable lifestyles and career pathways by transforming schools into living laboratories (125) is necessary to counter the often overlooked shadow side of education, since the secondary and higher levels of education are currently associated with higher resource use (126).

STIs through information feedbacks. The last tipping intervention is related to the flow of information and creating positive information feedbacks. The control parameter is represented by the transparency of the impact of individual consumer and lifestyle choices and carbon emissions. Transparency and disclosure of information about carbon emissions are needed, for instance, not just to provide a solid basis for global, regional, and national policies (127) but also to increase public and consumer awareness and improve labeling programs (128), triggering action and lifestyle changes to support decarbonization (129). The recent disclosure of the close ties between RWE, the biggest energy company in Germany, and regional politicians protecting their interest in the lignite coal extraction areas in Hessen led to a nationwide social movement and massive public demonstrations against plans to clear the Hambach Forest (130). Corporate disclosure of carbon assets can also help to overcome the short-term horizons of fund managers (131) and create a positive feedback in the divestment movement.

Another positive feedback can be identified between the information system and public education. Enhanced public knowledge and understanding by the broader public of the main variables and processes in the Earth's climate system and their linkages with human activities could increase public sensitivity to emissions-related information (132). Just as most product packages display nutritional facts, some authors propose they could display a second label on "Earth facts" and disclose the information on their carbon footprint and other emissions (133). In comparison, the global market for organic products, driven primarily by health concerns but clearly stimulated by providing clear labeling, increased at rates above 10% per year (134).

Discussion and Conclusions

Each of the STEs discussed above exists in the real world in varying degrees, locations, and scales and shows the potential to boost a decarbonization breakthrough. Since social-ecological dynamics are subject to complex processes that cannot be fully anticipated, it is not possible to predict when and where exactly tipping points will be crossed. However, the system can be imperfectly navigated intentionally to achieve certain desirable conditions and capacities (34). The social tipping dynamics are likely to spread through adaptive networks of interactions rather than via straightforward cause–effect systems. The identified interactions between the various STEs mean that they can potentially reinforce one another, making a transition to decarbonization more likely if several are triggered simultaneously (Fig. 3). In addition, crossing multiple tipping points in diverse systems of action increases the likelihood of breaking existing systemic inertia and lock-ins and thereby achieving the climate policy goals (34, 45). The interactions between the nominated candidates for STEs could be organized as different possible transformative pathways leading to crossing tipping points across scales and regions. These "tipping transformative pathways" can potentially show the bottom-up emergence of the global sustainability pathway (SSP1) (135).

One possible transformative pathway that has recently started to materialize has been initiated within the education system by school children who started the climate strikes #FridaysForFuture. The movement is causing "irritations" in personal worldviews (136) and thus might be changing peoples' norms and values and the ways of thinking and acting, possibly leading to changes in policies and regulations, infrastructure development, as well as individual consumption and lifestyle decisions. For example, as a result of the massive school student protest in Germany, even the traditionally climate-conservative parties recently started to address climate change issues in their programs (137). The increasing awareness of the seriousness of climate change might drive an increasing demand for greenhouse-gas emission disclosure of various products and services. It might also drive an increased recognition of the intergenerationally unethical and immoral character of fossil fuels that will furthermore strengthen the legitimacy of carbon mitigation policies, including the removal of fossil-fuel subsidies. Although changes of norms, customs, and beliefs occur very slowly (138), one should keep mind that now is not year zero of the global sustainability transformation. It has now been 30 y since the Intergovernmental Panel on Climate Change (IPCC) was endorsed by the United Nations and issued its first report recognizing the anthropogenic character of climate change, and many important milestones have been reached since then, including publishing the subsequent IPCC reports, Pope Francis's encyclical *Laudato si'*, and numerous events led by artists and activists increasing the concern about climate issues. The example of "flight shaming" that was initiated by a Swedish Olympic athlete and has been popularized in social media (139), shows that society may now be just at the edge of tipping in the realm of social norms and beliefs. The high number of seats that environmentally oriented parities recently won in the European Union (EU) elections (140) shows that EU policy might potentially undergo a substantial shift within the next few years, the EU becoming a global leader in carbon mitigation efforts.

A global breakthrough could also be initiated at the level of resource allocation by redirecting financial flows in line with the divestment movement and improving information feedbacks by disclosing the greenhouse gas emissions of products and services. At this level, firms take consumption and production decisions constrained by budget as well as by information and technology availability (20, 51). Changes at this level occur continuously. Very rapid changes, at a rate of 50% or more, can occur within a few months. This is shown by public opinion polls on, for example, political preferences following information flows, particularly in online social media (141). Rapid changes in stock markets can occur within hours (142). Nevertheless, such trends rarely lead to bigger changes in human societies without simultaneous institutional

Reproduced from: I. M. Otto et al., Social tipping dynamics for stabilizing Earth's climate by 2050, Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 5, pp. 2354–2365, 2020, doi: 10.1073/pnas.1900577117. Published under Creative Commons Attribution License 4.0 (CC BY).

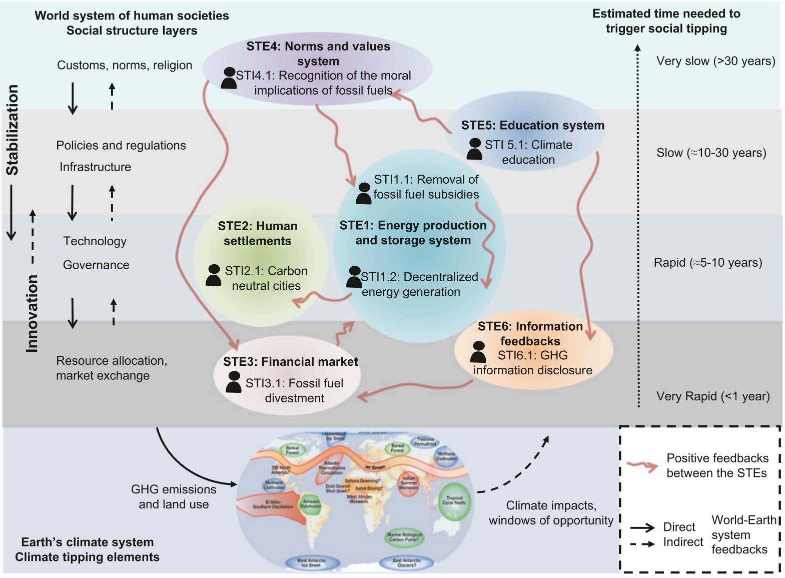

World system of human societies
Social structure layers
Estimated time needed to trigger social tipping
Customs, norms, religion
Policies and regulations
Infrastructure
Technology
Governance
Resource allocation, market exchange
Stabilization
Innovation
STE4: Norms and values system
STI4.1: Recognition of the moral implications of fossil fuels
STE5: Education system
STI 5.1: Climate education
STI1.1: Removal of fossil fuel subsidies
STE2: Human settlements
STI2.1: Carbon neutral cities
STE1: Energy production and storage system
STI1.2: Decentralized energy generation
STE6: Information feedbacks
STI6.1: GHG information disclosure
STE3: Financial market
STI3.1: Fossil fuel divestment
Very slow (>30 years)
Slow (≈10-30 years)
Rapid (≈5-10 years)
Very Rapid (<1 year)
GHG emissions and land use
Climate impacts, windows of opportunity
Earth's climate system
Climate tipping elements
Greenland Ice Sheet
Boreal Forest
Yedoma Permafrost
Boreal Forest
Methane Clathrates
Atlantic Thermohaline Circulation
SW North America?
Jet Stream
Indian Summer Monsoon
Methane Clathrates
Sahara Greening?
Dust Source Shut-down?
Sahel Drying?
El Niño-Southern Oscillation
Amazon Rainforest
West African Monsoon
Tropical Coral Reefs
Marine Biological Carbon Pump?
West Antarctic Ice Sheet
East Antarctic Glaciers?
Positive feedbacks between the STEs
Direct
Indirect
World-Earth system feedbacks

Fig. 3. Social tipping elements (STEs) and associated social tipping interventions (STIs) with the potential to drive rapid decarbonization in the World–Earth system. The processes they represent unfold across levels of social structure on widely different timescales, ranging from the fast dynamics of market exchanges and resource allocation on subannual timescales to the slow decadal- to centennial-scale changes on the level of customs, values, and social norms (51).

changes. The institutional changes, requiring more time, such as transforming the public subsidies and taxation systems, are needed to stabilize the new emerging system. Otherwise the system might become increasingly unstable, bouncing back and forth between the old and new social order, delaying the transformation. A well-documented example of such a phenomenon is the rebound effect (143, 144). Even the frequently quoted "successful" example of feed-in tariffs and German energy transition "Energiewende" to renewables, which used the rapid change in public opinion in the aftermath of the nuclear catastrophe in Fukushima in Japan in 2011, have recently faded away due to the lack of sufficiently sustained societal and policy support (145).

Many of the nominated candidates for STEs extend beyond achieving greenhouse gas reduction and can be potentially interlinked with achieving other global policy goals, such as the Sustainable Development Goals. Many of the interventions discussed above include a range of well-being and public health cobenefits (68). Solving the climate crisis could be a chance to redesign the global socioeconomic institutions toward achieving a more just and equitable future (146). Several authors point out that environmental catastrophes, including increased severity and frequency of climatic extremes, might act as "windows of opportunity" that give rise to uncertainty and confusion, which might in turn motivate actors to engage in reflective processes and take sharp breaks from the existing procedures and policies (147) (Fig. 3). However, although the opportunity for a revolutionary change might emerge due to external or environmental factors (148), it is important to actively work with the social complexities (Fig. 2) and the relevant key social actors (Tables 2 and 3), to increase public acceptance and support for the transformative changes to come. To ensure that climate-related social learning will take place, it is necessary to understand how changes of perceptions and awareness, motives, and interests of various actors take place and how institutional innovations occur (149).

We call on both social and natural sciences to engage more intensively in collaborative interdisciplinary research to understand rapid social transformations, STEs, and their interactions with tipping elements in the Earth system. Planetary social-ecological models and machine-learning techniques can help to explore the control parameters and critical thresholds in the trajectory of this World–Earth coevolutionary dynamics (43). We also encourage studies on the archetypes of social transformations (150) in different world regions as well as using insights and methods from the natural sciences to study the complexity of social systems. Both empirical studies and modeling exercises could also help to assess the distributional impacts of STIs and factors influencing their effectiveness. Our study presents a comprehensive empirical analysis of social tipping dynamics for global decarbonization. However, since our results were derived from an elicitation process involving small and nonrepresentative samples of experts, more research is needed to verify our findings and to provide more robust empirical evidence and data. Experts from the research sector and the Global North were overrepresented in our sample. Therefore special attention should be given to the expertise of low-carbon and sustainability practitioners as well as to providing more empirical material from the Global South. Finally, the urgency and complex character of climate change require transdisciplinarity and engagement with social movements, knowledge brokers, and change leaders (151). More research is needed on understanding the required social processes and the drivers and incentives for short-term engagement of diverse coalitions of action around concrete solutions and strategies at various governance levels (152).

Materials and Methods

The primary data collection tool was an online expert survey that was sent to over 1,000 international experts through a private message or addressed through mailing lists of organizations in the field of climate change and sustainability. A full list of all survey questions as well details on the research organization are provided in *SI Appendix*. The survey ran for 2.5 mo, and it was completed by 133 experts. In total, they suggested 207 candidates for STEs and interventions instrumental for decarbonization by 2050. A selected group of 17 experts were invited for a workshop that focused on choosing the top candidates for STEs. Finally, the coauthors carried out a literature review on the top candidates selected at the workshop, following the literature review guidelines.

Data Availability Statement. All data discussed in the paper will be made available to readers upon request.

Reproduced from: I. M. Otto et al., Social tipping dynamics for stabilizing Earth's climate by 2050, Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 5, pp. 2354–2365, 2020, doi: 10.1073/pnas.1900577117. Published under Creative Commons Attribution License 4.0 (CC BY).

ACKNOWLEDGMENTS. We acknowledge financial support by the Earth League's EarthDoc program, the Stordalen Foundation (http://pb.net/ Network), the Leibniz Association (DominoES), and the European Research Council (Earth Resilience in the Anthropocene). We thank James Greyson, Claudia Köhler, Sangjun Park, Anja Klein, Andrzej Ancygier, Frank Wechsung, Doyne Farmer, Tim Lenton, Jobst Heitzig, Ricarda Winkelmann, Alison Schlums, Maria Martin, Konrad Hagedorn, the WINS Group, and the research participants.

1. IPCC, *Global Warming of 1.5°C: An IPCC Special Report on the Impacts of Global Warming of 1.5°C above Pre-Industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*, V. Masson-Delmotte *et al*., Eds. (World Meteorological Organization, Geneva, Switzerland).
2. W. Steffen *et al*., Trajectories of the Earth system in the Anthropocene. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 8252–8259 (2018).
3. W. V. Reid *et al*., Environment and development. Earth system science for global sustainability: Grand challenges. *Science* **330**, 916–917 (2010).
4. W. Steffen *et al*., Sustainability. Planetary boundaries: Guiding human development on a changing planet. *Science* **347**, 1259855 (2015).
5. United Nations. Paris Agreement (2015). https://unfccc.int/process-and-meetings/the-paris-agreement/the-paris-agreement. Accessed 20 February 2017.
6. J. Rockström *et al*., A roadmap for rapid decarbonization. *Science* **355**, 1269–1271 (2017).
7. F. W. Geels, Regime resistance against low-carbon transitions: Introducing politics and power into the multi-level perspective. *Theory Cult. Soc.* **31**, 21–40 (2014).
8. C. Kuzemko, M. Lockwood, C. Mitchell, R. Hoggett, Governing for sustainable energy system change: Politics, contexts and contingency. *Energy Res. Soc. Sci.* **12**, 96–105 (2016).
9. B. P. Heard, B. W. Brook, T. M. L. Wigley, C. J. A. Bradshaw, Burden of proof: A comprehensive review of the feasibility of 100% renewable-electricity systems. *Renew. Sustain. Energy Rev.* **76**, 1122–1133 (2017).
10. H.-W. Sinn, Buffering volatility: A study on the limits of Germany's energy revolution. *Eur. Econ. Rev.* **99**, 130–150 (2017).
11. World Bank and Ecofys, *Carbon Pricing Watch 2017* (World Bank Group, 2017).
12. S. E. Shmelev, S. U. Speck, Green fiscal reform in Sweden: Econometric assessment of the carbon and energy taxation scheme. *Renew. Sustain. Energy Rev.* **90**, 969–981 (2018).
13. E. Tvinnereim, M. Mehling, Carbon pricing and deep decarbonisation. *Energy Policy* **121**, 185–189 (2018).
14. M. M. Holland, C. M. Bitz, B. Tremblay, Future abrupt reductions in the summer Arctic sea ice. *Geophys. Res. Lett.* **33**, L23503 (2006).
15. T. M. Lenton *et al*., Tipping elements in the Earth's climate system. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 1786–1793 (2008).
16. V. Dakos, S. R. Carpenter, E. H. van Nes, M. Scheffer, Resilience indicators: Prospects and limitations for early warnings of regime shifts. *Phil. Trans. R. Soc. B.* **370**, 20130263 (2015).
17. M. Grodzin, Metropolitan segregation. *Sci. Am.* **197**, 33–41 (1957).
18. T. C. Schelling, Dynamic models of segregation. *J. Math. Sociol.* **1**, 143–186 (1971).
19. C. Doyle, S. Sreenivasan, B. K. Szymanski, G. Korniss, Social consensus and tipping points with opinion inertia. *Phys. Stat. Mech. Its Appl.* **443**, 316–323 (2016).
20. J. D. Farmer *et al*., Sensitive intervention points in the post-carbon transition. *Science* **364**, 132–134 (2019).
21. B. Walker *et al*., A handful of heuristics and some propositions for understanding resilience in social-ecological systems. *Ecol. Soc.* **11**, 13 (2006).
22. C. Folke, T. Hahn, P. Olsson, J. Norberg, Adaptive governance of social-ecological systems. *Annu. Rev. Environ. Resour.* **30**, 441–473 (2005).
23. M. Milkoreit *et al*., Defining tipping points for social-ecological systems scholarship—an interdisciplinary literature review. *Environ. Res. Lett.* **13**, 033005 (2018).
24. M. U. J. Edwards, *Printing, Propaganda and Martin Luther* (Fortress Press, 2005).
25. J. E. Trancik *et al*., Technology improvement and emissions reductions as mutually reinforcing efforts. http://energy.mit.edu/publication/technology-improvement-and-emissions-reductions-as-mutually-reinforcing-efforts/. Accessed 20 August 2019.
26. G. Kavlak, J. McNerney, J. E. Trancik, Evaluating the causes of cost reduction in photovoltaic modules. *Energy Policy* **123**, 700–710 (2018).
27. H. J. Schellnhuber, Tipping elements in the Earth system. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 20561–20563 (2009).
28. R. E. Kopp, R. Shwom, G. Wagner, J. Yuan, Tipping elements and climate-economic shocks. *Earths Futur.*, 10.1002/2016EF000362 (2016).
29. E. Kriegler *et al*., A new scenario framework for climate change research: The concept of shared climate policy assumptions. *Clim. Change* **122**, 401–414 (2014).
30. A. Grubler *et al*., A low energy demand scenario for meeting the 1.5 °C target and Sustainable Development Goals without negative emission technologies. *Nat. Energy* **3**, 515–527 (2018).
31. R. C. Hilborn, Sea gulls, butterflies, and grasshoppers: A brief history of the butterfly effect in nonlinear dynamics. *Am. J. Phys.* **72**, 425–427 (2004).
32. E. N. Lorenz, Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141 (1963).
33. D. B. Bonar, Molluscan metamorphosis: A study in tissue transformation. *Integr. Comp. Biol.* **16**, 573–591 (1976).
34. J. D. Tàbara *et al*., Positive tipping points in a rapidly warming world. *Curr. Opin. Environ. Sustain.* **31**, 120–129 (2018).
35. S. Lehmann, Y.-Y. Ahn, "Spreading in social systems: Reflections" in *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*, S. Lehmann, Y.-Y. Ahn, Eds. (Springer International Publishing, 2018), pp. 351–358.
36. D. Centola, J. Becker, D. Brackbill, A. Baronchelli, Experimental evidence for tipping points in social convention. *Science* **360**, 1116–1119 (2018).
37. C.-F. Schleussner *et al*., Differential climate impacts for policy-relevant limits to global warming: The case of 1.5 °C and 2 °C. *Earth Syst. Dyn.* **7**, 327–351 (2016).
38. K. Nyborg *et al*., Social norms as solutions. *Science* **354**, 42–43 (2016).
39. C. Loftin, Assaultive violence as a contagious social process. *Bull. N. Y. Acad. Med.* **62**, 550–555 (1986).
40. P. Hedström, Contagious collectivities: On the spatial diffusion of Swedish trade unions, 1890–1940. *Am. J. Sociol.* **99**, 1157–1179 (1994).
41. N. A. Christakis, J. H. Fowler, The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.* **357**, 370–379 (2007).
42. N. A. Christakis, J. H. Fowler, The collective dynamics of smoking in a large social network. *N. Engl. J. Med.* **358**, 2249–2258 (2008).
43. J. F. Donges *et al*., Earth system modelling with complex dynamic human societies: The copan:CORE World-Earth modeling framework. https://doi.org/10.5194/esd-2017-126 (15 January 2018).
44. G. Spaargaren, *The Ecological Modernization of Production and Consumption. Essays in Environmental Sociology* (University of Wageningen, 1997).
45. I. M. Otto *et al*., Human agency in the anthropocene. *Ecol. Econ.* **167**, 106463 (2020).
46. J. D. Tàbara, C. Pahl-Wostl, Sustainability learning in natural resource use and management. *Ecol. Soc.* **12**, 3 (2007).
47. G. Schreyögg, J. Sydow, Organizational path dependence: A process view. *Organ. Stud.* **32**, 321–335 (2011).
48. E. Dimant, Contagion of pro- and anti-social behavior among peers and the role of social proximity. *J. Econ. Psychol.* **73**, 66–88 (2019).
49. P. Sztompka, Cultural trauma: The other face of social change. *Eur. J. Soc. Theory* **3**, 449–466 (2000).
50. United Nations, UN Sustainable Development Goals (2015). https://sustainabledevelopment.un.org/?menu=1300. Accessed 20 May 2017.
51. O. E. Williamson, The new institutional economics: Taking stock, looking ahead. *J. Econ. Lit.* **38**, 595–613 (2000).
52. C. Rammel, S. Stagl, H. Wilfing, Managing complex adaptive systems—a co-evolutionary perspective on natural resource management. *Ecol. Econ.* **63**, 9–21 (2007).
53. R. Koch, *The 80/20 Principle: The Secret to Achieving More with Less* (Crown Publishing Group, 2011).
54. M. Gladwell, *The Tipping Point: How Little Things Can Make a Big Difference* (Back Bay Books, 2000).
55. IPCC, *Climate Change 2014: Mitigation of Climate Change* (Cambridge University Press, 2015).
56. International Renewable Energy Agency. REthinking Energy 2017: Accelerating the global energy transformation (International Renewable Energy Agency, Abu Dhabi, United Arab Emirates, 2017).
57. Breakthrough Energy, Breakthrough Energy Announces Expanded Coalition and Fund Focus Areas, 12 December 2017. https://www.b-t.energy/media-resources/media-release-dec-2017/. Accessed 20 February 2018.
58. A. D. Foster, M. R. Rosenzweig, Microeconomics of technology adoption. *Annu. Rev. Econ.* **2**, 395–424 (2010).
59. A. McCrone, U. Moslener, F. d'Estais, C. Grüning, Global trends in renewable energy investments. https://euagenda.eu/upload/publications/untitled-87074-ea.pdf. Accessed 20 April 2018.
60. G. F. Nemet, Beyond the learning curve: Factors influencing cost reductions in photovoltaics. *Energy Policy* **34**, 3218–3232 (2006).
61. G. J. Dalton, D. A. Lockington, T. E. Baldock, Feasibility analysis of renewable energy supply options for a grid-connected large hotel. *Renew. Energy* **34**, 955–964 (2009).
62. M. Ram *et al*., *Global Energy System Based on 100% Renewable Energy—Power, Heat, Transport and Desalination Sectors* (Lappeenranta University of Technology and Energy Watch Group, 2019).
63. D. L. B. Schwappach, Resource allocation, social values and the QALY: A review of the debate and empirical evidence. *Health Expect.* **5**, 210–222 (2002).
64. P. Griffin, *The Carbon Majors Database CDP: Carbon Majors Report 2017*. https://b8f65cb373b1b7b15feb-c70d8ead6ced550b4d987d7c03fcdd1d.ssl.cf3.rackcdn.com/cms/reports/documents/000/002/327/original/Carbon-Majors-Report-2017.pdf. Accessed 20 February 2018.
65. International Energy Agency, *World Energy Outlook 2010*. https://webstore.iea.org/world-energy-outlook-2010. Accessed 18 April 2017.
66. H. van Asselt, K. Kulovesi, Seizing the opportunity: Tackling fossil fuel subsidies under the UNFCCC. *Int. Environ. Agreement Polit. Law Econ.* **17**, 357–370 (2017).
67. M. Jakob, C. Chen, S. Fuss, A. Marxen, O. Edenhofer, Development incentives for fossil fuel subsidy reform. *Nat. Clim. Chang.* **5**, 709–712 (2015).
68. D. Coady, I. Parry, L. Sears, B. Shang, How large are global energy subsidies? https://www.imf.org/en/Publications/WP/Issues/2016/12/31/How-Large-Are-Global-Energy-Subsidies-42940. Accessed 20 July 2019.
69. R. McKenna, The double-edged sword of decentralized energy autonomy. *Energy Policy* **113**, 747–750 (2018).
70. M. L. Di Silvestre, S. Favuzza, E. Riva Sanseverino, G. Zizzo, How decarbonization, digitalization and decentralization are changing key power infrastructures. *Renew. Sustain. Energy Rev.* **93**, 483–498 (2018).
71. P. Denholm, M. Hand, Grid flexibility and storage required to achieve very high penetration of variable renewable electricity. *Energy Policy* **39**, 1817–1830 (2011).

Reproduced from: I. M. Otto et al., Social tipping dynamics for stabilizing Earth's climate by 2050, Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 5, pp. 2354–2365, 2020, doi: 10.1073/pnas.1900577117. Published under Creative Commons Attribution License 4.0 (CC BY).

72. J. A. P. Lopes, N. Hatziargyriou, J. Mutale, P. Djapic, N. Jenkins, Integrating distributed generation into electric power systems: A review of drivers, challenges and opportunities. *Electr. Power Syst. Res.* **77**, 1189–1203 (2007).
73. M. Wolsink, The research agenda on social acceptance of distributed generation in smart grids: Renewable as common pool resources. *Renew. Sustain. Energy Rev.* **16**, 822–835 (2012).
74. D. MacKay, *Sustainable Energy—Without the Hot Air* (UIT Cambridge, 2008).
75. D. Helm, *The Carbon Crunch: Revised and Updated* (Yale University Press, 2015).
76. R. J. Hewitt *et al*., Social innovation in community energy in Europe: A review of the evidence. *Front. Energy Res.* **7**, 31 (2019).
77. L. Roy, REWDT trials new Heat Smart project on household heating devices in Orkney (2016). https://www.power-technology.com/uncategorised/newsrewdt-trials-heat-smart-project-household-heating-devices-orkney-4957725/. Accessed 20 July 2019.
78. N. Magnani, G. Osti, Does civil society matter? Challenges and strategies of grassroots initiatives in Italy's energy transition. *Energy Res. Soc. Sci.* **13**, 148–157 (2016).
79. T. P. Wright, Factors affecting the cost of airplanes. *J. Aeronaut. Sci.* **3**, 122–128 (1936).
80. P. M. McGuirk, H. Bulkeley, R. Dowling, Configuring urban carbon governance: Insights from Sydney, Australia. *Ann. Am. Assoc. Geogr.* **106**, 145–166 (2016).
81. A. Aylett, Networked urban climate governance. *Environ. Plann. C Gov. Policy* **31**, 858–875 (2013).
82. S. J. Davis *et al*., Net-zero emissions energy systems. *Science* **360**, eaas9793 (2018).
83. P. Hernandez, P. Kenny, From net energy to zero energy buildings: Defining life cycle zero energy buildings (LC-ZEB). *Energy Build.* **42**, 815–821 (2010).
84. J. Laukkonen *et al*., Combining climate change adaptation and mitigation measures at the local level. *Habitat Int.* **33**, 287–292 (2009).
85. D. M. Roodman, N. Lenssen, *A Building Revolution: How Ecology and Health Concerns are Transforming Construction* (Worldwatch Institute, 1995).
86. M. E. Porter, Location, competition, and economic development: Local clusters in a global economy. *Econ. Dev. Q.* **14**, 15–34 (2000).
87. J. W. G. Van De Kuilen, A. Cecotti, Z. Xia, M. He, Very tall wooden buildings with cross laminated timber. *Procedia Eng.* **14**, 1621–1628 (2011).
88. N. Bardsley, J. Mehta, C. Starmer, R. Sugden, Explaining focal points: Cognitive hierarchy theory versus team reasoning. *Econ. J. (Lond.)* **120**, 40–79 (2010).
89. G. Feola, R. Nunes, Success and failure of grassroots innovations for addressing climate change: The case of the Transition Movement. *Glob. Environ. Change* **24**, 232–250 (2014).
90. Energy Cities, Vision and mission. https://energy-cities.eu/vision-mission/. Accessed 30 December 2019.
91. Energy Cities, Fossil fuel free Växjö. http://enercitee.eu/files/dokumente/Policy_maker_exchange/Fossil_Fuel_Free_Vaxjo_-_the_story_2010.pdf. Accessed 20 May 2018.
92. M. Campello, J. R. Graham, C. R. Harvey, The real effects of financial constraints: Evidence from a financial crisis. *J. Financ. Econ.* **97**, 470–487 (2010).
93. B. Ott, "The carbon asset bubble—mythos or menace? An attempted refutation" Bachelor thesis, Hochschule für Wirtschaft und Recht Berlin, Berlin, Germany (2016).
94. J. Ritchie, H. Dowlatabadi, Divest from the carbon bubble? Reviewing the implications and limitations of fossil fuel divestment for institutional investors. *Rev. Econ. Finance* **5**, 59–80 (2015).
95. J. Rubin, *The Carbon Bubble* (Random House, 2015).
96. B. Ewers, J. F. Donges, J. Heitzig, S. Peterson, Divestment may burst the carbon bubble if investors' beliefs tip to anticipating strong future climate policy. arXiv: 1902.07481 (20 February 2019).
97. G. A. Lenferna, "Divestment as climate justice: Weighing the power of the fossil fuel divestment movement" in *Climate Justice and the Economy: Social Mobilization, Knowledge and the Political*, S. G. Jacobsen, Ed. (Routledge Earth Scan, 2018), pp. 84–109.
98. D. Carrington, Fossil fuel divestment funds double to $5tn in a year, *The Guardian*, 16 December 2016. https://www.theguardian.com/environment/2016/dec/12/fossil-fuel-divestment-funds-double-5tn-in-a-year. Accessed 20 September 2018.
99. N. Healy, J. Barry, Politicizing energy justice and energy system transitions: Fossil fuel divestment and a "just transition." *Energy Policy* **108**, 451–459 (2017).
100. O. Ralph, Insurers go cold on coal industry. *Financial Times*, 7 January 2018. https://www.ft.com/content/7ec63f34-f20c-11e7-ac08-07c3086a2625. Accessed 20 May 2018.
101. A. Vaughan, World's biggest sovereign wealth fund proposes ditching oil and gas holdings. *The Guardian*, 16 November 2017. https://www.theguardian.com/business/2017/nov/16/oil-and-gas-shares-dip-as-norways-central-bank-advises-oslo-to-divest. Accessed 22 May 2018.
102. I. M. Otto *et al*., Social vulnerability to climate change: A review of concepts and evidence. *Reg. Environ. Change* **17**, 1651–1662 (2017).
103. H. J. Schellnhuber *et al*., "The challenge of a 4 degrees celsius world by 2100" in *Hexagon Series on Human Environmental Security and Peace*, H. G. Brauch, Ed. (Springer, 2016), pp. 267–283.
104. M. Kampa, E. Castanas, Human health effects of air pollution. *Environ. Pollut.* **151**, 362–367 (2008).
105. P. Simpson, C. Hill, "Leadership, spirituality and complexity: Wilberforce and the abolition of the slave trade" in *Leadership Perspectives*, K. T. James, J. Collins, Eds. (Palgrave Macmillan, 2008), pp. 29–42.
106. R. Gifford, The dragons of inaction: Psychological barriers that limit climate change mitigation and adaptation. *Am. Psychol.* **66**, 290–302 (2011).
107. E. M. Rogers, *Diffusion of Innovations* (Simon and Schuster, ed. 4, 2010).
108. P. Francis, *Encyclical Letter Laudato Si' of the Holy Father Francis on Care for Our Common Home* (Vatikan Press, 2015).
109. C. Freund, S. Oliver, The Origins of the superrich: The billionaire characteristics database (2016). https://papers.ssrn.com/abstract=2731353. Accessed 6 February 2019.
110. R. A. Posner, Social norms and the law: An economic approach. *Am. Econ. Rev.* **87**, 365–369 (1997).
111. F. Green, Anti-fossil fuel norms. *Clim. Change* **150**, 103–116 (2018).
112. G. J. M. Velders, S. O. Andersen, J. S. Daniel, D. W. Fahey, M. McFarland, The importance of the Montreal Protocol in protecting climate. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 4814–4819 (2007).
113. D. Evensen, The rhetorical limitations of the #FridaysForFuture movement. *Nat. Clim. Chang.* **9**, 428–433 (2019).
114. J. Lin, *Social Transformation and Private Education in China* (Praeger Publishers, 1999).
115. W. Lutz, R. Muttarak, E. Striessnig, Environment and development. Universal education is key to enhanced climate adaptation. *Science* **346**, 1061–1062 (2014).
116. T. M. Lee, E. M. Markowitz, P. D. Howe, C.-Y. Ko, A. A. Leiserowitz, Predictors of public climate change awareness and risk perception around the world. *Nat. Clim. Chang.* **5**, 1014–1020 (2015).
117. E. Plutzer *et al*., Climate confusion among U.S. teachers. *Science* **351**, 664–665 (2016).
118. I. Lorenzoni, S. Nicholson-Cole, L. Whitmarsh, Barriers perceived to engaging with climate change among the UK public and their policy implications. *Glob. Environ. Change* **17**, 445–459 (2007).
119. Z. W. Kundzewicz *et al*., Uncertainty in climate change impacts on water resources. *Environ. Sci. Policy* **79**, 1–8 (2018).
120. A. Kempf, The Cuban literacy campaign at 50: Formal and tacit learning in revolutionary education. *Crit. Educ.*, 10.14288/ce.v5i4.183269 (2014).
121. S. Suranovic, Fossil fuel addiction and the implications for climate change policy. *Glob. Environ. Change* **23**, 598–608 (2013).
122. K. E. Warner, The effects of the anti-smoking campaign on cigarette consumption. *Am. J. Public Health* **67**, 645–650 (1977).
123. T. Dietz, G. T. Gardner, J. Gilligan, P. C. Stern, M. P. Vandenbergh, Household actions can provide a behavioral wedge to rapidly reduce US carbon emissions. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 18452–18456 (2009).
124. Life, *R.A.C.E.S. (Raising Awareness on Climate Change and Energy Saving): Final Report*. http://ec.europa.eu/environment/life/project/Projects/index.cfm?fuseaction=home.showFile&rep=file&fil=LIFE07_INF_IT_000487_FTR.pdf. Accessed 20 April 2018.
125. M. S. McCaffrey, *Climate Smart & Energy Wise: Advancing Science Literacy, Knowledge, and Know-How* (Corwin, 2014).
126. D. W. O'Neill, A. L. Fanning, W. F. Lamb, J. K. Steinberger, A good life for all within planetary boundaries. *Nat. Sustain.* **1**, 88–95 (2018).
127. M. Freedman, J. D. Park, A. J. Stagliano, "Mandated climate change disclosures: A study of large US firms that emit carbon dioxide" in *Sustainability and Governance*, C. R. Lehman, Ed. (Emerald Publishing, 2015), pp. 99–121.
128. D. Brounen, N. Kok, On the economics of energy labels in the housing market. *J. Environ. Econ. Manage.* **62**, 166–179 (2011).
129. J. Andrew, C. Cortese, Accounting for climate change and the self-regulation of carbon disclosures. *Account. Forum* **35**, 130–138 (2011).
130. H. Forst, Das ist die Mitte der Gesellschaft. *Zeit*, 6 October 2018. https://www.zeit.de/gesellschaft/2018-10/hambacher-forst-demonstration-rodung-rwe-armin-laschet. Accessed 24 October 2018.
131. J. Ayling, N. Gunningham, Non-state governance and climate policy: The fossil fuel divestment movement. *Clim. Policy* **17**, 131–149 (2017).
132. I. Otto-Banaszak, P. Matczak, J. Wesseler, F. Wechsung, Different perceptions of adaptation to climate change: A mental model approach applied to the evidence from expert interviews. *Reg. Environ. Change* **11**, 217–228 (2011).
133. K. Meyer, P. Newman, The planetary accounting framework: A novel, quota-based approach to understanding the impacts of any scale of human activity in the context of the planetary boundaries. *Sustain. Earth* **1**, 4 (2018).
134. A. Sahota, "The global market for organic food and drink" in *The World of Organic Agriculture: Statistics and Emerging Trends 2008*, H. Willer, M. Yussefi-Menzler, N. Sorensen, Eds. (Routledge, 2008), pp. 53–58.
135. B. C. O'Neill *et al*., A new scenario framework for climate change research: The concept of shared socioeconomic pathways. *Clim. Change* **122**, 387–400 (2014).
136. J. Mezirow, Transformative learning: Theory to practice. *New Dir. Adult Contin. Educ.* **1997**, 5–12 (1997).
137. F. Gathmann, C. Hecking, *CDU und CSU im Ökomodus: Plötzlich grün* (Spiegel Online, 2019).
138. O. E. Williamson, Transaction costs economics: How it works; where it is headed. *Econ.* **146**, 23–58 (1998).
139. H. Coffey, What is "flygskam"? Everything you need to know about the environmental movement that's sweeping Europe. *The Independent*, 5 June 2019. https://www.independent.co.uk/travel/news-and-advice/flygskam-anti-flying-flight-shaming-sweden-greta-thornberg-environment-air-travel-train-brag-a8945196.html. Accessed 7 June 2019.
140. E. Graham-Harrison, A quiet revolution sweeps Europe as Greens become a political force. *The Observer*, 2 June 2019. https://www.theguardian.com/politics/2019/jun/02/european-parliament-election-green-parties-success. Accessed 7 June 2019.
141. B. O'Connor, R. Balasubramanyan, B. R. Routledge, N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series" in *Proceedings of the International AAAI Conference on Weblogs and Social Media* (Association for the Advancement of Artificial Intelligence, 2010), pp. 122–129.
142. D. M. Kotz, The financial and economic crisis of 2008: A systemic crisis of neoliberal capitalism. *Rev. Radic. Polit. Econ.* **41**, 305–317 (2009).
143. W. S. Jevons, *The Coal Question: An Inquiry Concerning the Progress of the Nation, and the Probable Exhaustion of Our Coal-Mines* (Augustus M. Kelley, 1965).
144. R. Cremades *et al*., Co-benefits and trade-offs in the water–energy nexus of irrigation modernization in China. *Environ. Res. Lett.* **11**, 054007 (2016).
145. C. Kemfert, Germany must go back to its low-carbon future. *Nature* **549**, 26–27 (2017).

Reproduced from: I. M. Otto et al., Social tipping dynamics for stabilizing Earth's climate by 2050, Proc. Natl. Acad. Sci. U. S. A., vol. 117, no. 5, pp. 2354–2365, 2020, doi: 10.1073/pnas.1900577117. Published under Creative Commons Attribution License 4.0 (CC BY).

146. WBGU, *Welt im Wandel. Gesellschaftsvertrag für eine Große Tranformation* (Wissenschaftlicher Beitrat der Budesregierung Globale Umweltveränderungen, 2011).
147. T. R. Burns, "Two conceptions of human agency: Rational choice theory and the social theory of action" in *Agency and Structure. Reorienting Social Theory*, Piotr Sztompka, Ed. (Gordon and Breach Science Publishers, 1994), pp. 197–250.
148. W. J. Abernathy, K. B. Clark, Innovation: Mapping the winds of creative destruction. *Res. Policy* **14**, 3–22 (1985).
149. J. D. Tàbara *et al*., The climate learning ladder. A pragmatic procedure to support climate adaptation. *Environ. Policy Gov.* **20**, 1–11 (2010).
150. K. Eisenack *et al*., Design and quality criteria for archetype analysis. *Ecol. Soc.* **24**, 6 (2019).
151. F. Westley *et al*., Tipping toward sustainability: Emerging pathways of transformation. *Ambio* **40**, 762–780 (2011).
152. S. Aakre, S. Kallbekken, R. V. Dingenen, D. G. Victor, Incentives for small clubs of Arctic countries to limit black carbon and methane emissions. *Nat. Clim. Chang.* **8**, 85 (2018).
153. D. P. van Vuuren *et al*., The representative concentration pathways: An overview. *Clim. Change* **109**, 5–31 (2011).
154. J. Rogelj, M. Meinshausen, R. Knutti, Global warming under old and new scenarios using IPCC climate sensitivity range estimates. *Nat. Clim. Chang.* **2**, 248–253 (2012).
155. M. Scheffer *et al*., Early-warning signals for critical transitions. *Nature* **461**, 53–59 (2009).
156. A. Ganopolski, S. Rahmstorf, Abrupt glacial climate changes due to stochastic resonance. *Phys. Rev. Lett.* **88**, 038501 (2002).
157. J. H. Jo, J. S. Golden, S. W. Shin, Incorporating built environment factors into climate change mitigation strategies for Seoul, South Korea: A sustainable urban systems framework. *Habitat Int.* **33**, 267–275 (2009).
158. J. Lipp, Lessons for effective renewable electricity policy from Denmark, Germany and the United Kingdom. *Energy Policy* **35**, 5481–5495 (2007).
159. M. A. Boutabba, S. Lardic, EU emissions trading scheme, competitiveness and carbon leakage: New evidence from cement and steel industries. *Ann. Oper. Res.* **255**, 47–61 (2017).
160. D. W. Keith, G. Wagner, C. L. Zabel, Solar geoengineering reduces atmospheric carbon burden. *Nat. Clim. Chang.* **7**, 617–619 (2017).
161. S. Jacobsson, V. Lauber, The politics and policy of energy system transformation—Explaining the German diffusion of renewable energy technology. *Energy Policy* **34**, 256–276 (2006).
162. O. Wagner, K. Berlo, "The wave of remunicipalisation of energy networks and supply in Germany : The establishment of 72 new municipal power utilities" in *ECEEE 2015 Summer Study on Energy Efficiency: First Fuel Now* (European Council for an Energy Efficient Economy, Stockholm, 2015).
163. R. Bolton, T. J. Foxon, Infrastructure transformation as a socio-technical process—Implications for the governance of energy distribution networks in the UK. *Technol. Forecast. Soci. Change* **90**, 538–550 (2015).
164. B. Zeng *et al*., Integrated planning for transition to low-carbon distribution system with renewable energy generation and demand response. *IEEE Trans. Power Syst.* **29**, 1153–1165 (2014).
165. A. Yadoo, H. Cruickshank, The role for low carbon electrification technologies in poverty reduction and climate change strategies: A focus on renewable energy mini-grids with case studies in Nepal, Peru and Kenya. *Energy Policy* **42** (suppl. C), 591–602 (2012).
166. D. Hoornweg, L. Sugar, and C. L. T. Gómez, Cities and greenhouse gas emissions: Moving forward. *Environ. Urban.* **23**, 207–227 (2011).
167. J. Ritchie, H. Dowlatabadi, Life cycle impacts of divestment: Applying an economic input-output LCA model to measure financed emissions. https://open.library.ubc.ca/cIRcle/collections/graduateresearch/42591/items/1.0075807. Accessed 28 September 2018.
168. A. Hares, J. Dickinson, K. Wilkes, Climate change and the air travel decisions of UK tourists." *J. Transp. Geogr.* **18**, 466–473 (2010).
169. The Lancet Planetary Health, Power to the children. *Lancet Planet. Health* **3**, PE102 (2019).
170. P. C. Stern, T. Dietz, T. Abel, G. A. Guagnano, L. Kalof, A value-belief-norm theory of support for social movements: The case of environmentalism. *Hum. Ecol. Rev.* **6**, 81–97 (1999).
171. A. M. Padilla, W. Perez, Acculturation, social identity, and social cognition: A new perspective. *Hisp. J. Behav. Sci.* **25**, 35–55 (2003).
172. E. A. Nadelmann, Global prohibition regimes: The evolution of norms in international society. *Int. Organ.* **44**, 479–526 (1990).
173. M. Story, M. S. Nanney, M. B. Schwartz, Schools and obesity prevention: Creating school environments and policies to promote healthy eating and physical activity. *Milbank Q.* **87**, 71–100 (2009).
174. S. Cowan, C. Deegan, Corporate disclosure reactions to Australia's first national emission reporting scheme. *Accounting & Finance* **51**, 409–436 (2011).
175. P. Upham, L. Dendler, M. Bleda, Carbon labelling of grocery products: Public perceptions and potential emissions reductions. *J. Clean. Prod.* **19**, 348–355 (2011).
176. A. Fraser, Are investment carbon footprints good for investors and the climate? *Policy Options* (2017). http://policyoptions.irpp.org/magazines/november-2017/are-investment-carbon-footprints-good-for-investors-and-the-climate/. Accessed 20 September 2018.
177. A. Banerjee, B. D. Solomon, Eco-labeling for energy efficiency and sustainability: A meta-evaluation of US programs. *Energy Policy* **31**, 109–123 (2003).
178. I. Siró, E. Kápolna, B. Kápolna, A. Lugasi, Functional food. Product development, marketing and consumer acceptance—A review. *Appetite* **51**, 456–467 (2008).

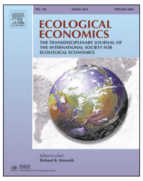
Ecological Economics 167 (2020) 106463

Contents lists available at ScienceDirect

Ecological Economics

journal homepage: www.elsevier.com/locate/ecolecon

ELSEVIER



Analysis

Human agency in the Anthropocene

Ilona M. Otto[a,*], Marc Wiedermann[a], Roger Cremades[b], Jonathan F. Donges[a,d], Cornelia Auer[a], Wolfgang Lucht[a,c]

[a] *Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany*
[b] *Climate Service Center Germany (GERICS), Hamburg, Germany*
[c] *Department of Geography, Humboldt University, Berlin, Germany*
[d] *Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden*

ARTICLE INFO

Keywords:
Human agency
Anthropocene
Integrated assessment models
Socio-metabolic agent classes

ABSTRACT

The human species has been recognized as a new force that has pushed the Earth's system into a new geological epoch referred to as the Anthropocene. This human influence was not conscious, however, but an unintended effect of the consumption of fossil-fuels over the last 150 years. Do we, humans, have the agency to deliberately influence the fate of our species and the planet we inhabit? The rational choice paradigm that dominated social sciences in the 20th Century, and has heavily influenced the conceptualization of human societies in global human-environmental system modelling in the early 21st Century, suggests a very limited view of human agency. Humans seen as rational agents, coordinated through market forces, have only a very weak influence on the system rules. In this article we explore alternative concepts of human agency that emphasize its collective and strategic dimensions as well as we ask how human agency is distributed within the society. We also explore the concept of social structure as a manifestation of, and a constraint on, human agency. We discuss the implications for conceptualization of human agency in integrated assessment modelling efforts.

1. Introduction

The Sustainable Development Goals and the Paris Agreement set very ambitious goals that, if taken seriously, would result in a rapid transformation of human-environmental interactions and decarbonization of the global socio-economic system (United Nations, 2015a, 2015b). What the agreements do not specify, however, is how the transformation should be achieved and who the transformation agents would be. In most modern scientific assessment of global human-environmental interactions, including Integrated Assessment Models (IAMs), alternative futures do not evolve from the behavior of the population in the simulated region or market, but are externally chosen by the research teams (e.g. Moss et al., 2010). The human agency that can be broadly understood as the capacity of individual and collective actors to change the course of events or the outcome of processes (Pattberg and Stripple, 2008) is only weakly represented in the commonly used global system models. For example, Integrated Assessment Models are not capable of modelling abrupt changes and tipping points in both natural and human systems (e.g. van Vuuren et al., 2012) that may imply severe and non-linear consequences for the Earth system as a whole (Lenton et al., 2008). There is, however, a relatively rich body of literature in social sciences, primarily in political science and institutional theory, that conceptualizes human agency in the governance of social-ecological systems (e.g. Ostrom, 2005; Kashwan et al., 2018) and in Earth system governance (e.g. Biermann et al., 2012, 2016). The aim of this paper is to assess the representation of human agency in Earth system science and integrated assessment modelling efforts and to examine how the rich body of literature on human agency in social sciences could be used to improve the modelling efforts.

The cornerstones of social sciences are built on the tension between agency and structure in social reproduction - the force of self-determination versus the embeddedness of social institutions (Dobres and Robb, 2000). Just as bio-physical laws determine the coupling between chemical and mechanical processes, social structures, including norms and institutions, impose constraints on the shaping of human interactions (North, 1990); they specify what people may, must, or must not do under particular circumstances and impose costs for non-compliance (Ostrom, 2005). Social institutions also have a function in expressing common or social interest and in channeling human behavior into what is socially desired (Coleman, 1990). Unlike bio-physical laws, however, social institutions are man-made structures and they are constantly being transformed by human action. In general, the smaller the social

[*] Corresponding author.
E-mail address: ilona.otto@pik-potsdam.de (I.M. Otto).

https://doi.org/10.1016/j.ecolecon.2019.106463
Received 5 September 2018; Received in revised form 31 August 2019; Accepted 9 September 2019

Reproduced from: I. M. Otto et al., Human agency in the Anthropocene, Ecol. Econ., vol. 167, 2020, doi: 10.1016/j.ecolecon.2019.106463, with the permission of Elsevier

entity the less durable it is. The size, scale, and time-frame of the social entity push it towards a durable structure and stability (Fuchs, 2001). Numerous authors have contributed to this long and fruitful debate on micro- and macro-level social structures and interactions within social sciences. However, very little of that knowledge has so far been applied by the global environmental change modelling community. To give an example, the IPCC Report on Mitigation of Climate Change underlines the role of institutional, legal, and cultural barriers that constrain the low-carbon technology uptake and behavioral change. However, the diffusion of alternative values, institutions, and even technologies are not incorporated in the modelling results (Edenhofer et al., 2014). Little is known about the potential for scaling-up of social innovations, let alone the possible carbon emission reductions they could drive if applied on a larger scale. How quickly would such innovations diffuse into virtual and face-to-face social networks, and what would the agency of different actors, and groups of actors, be in such a diffusion process? The purpose of this work is to analyze how social theory could be better integrated into the global environmental change assessment community, and how relevant social theory could be incorporated in modelling efforts.

The paper is structured as follows. We start by reviewing how human agency has been incorporated within Earth system science and integrated modelling efforts so far. We then move to the exploration of the concept of human agency and social structure and review the relevant social stratification theories. We propose how the concept of human agency could be incorporated in global human-environmental system models, and finally we conclude.

2. Human agency in Earth system science and integrated assessment modelling

The recognition of the human species as the driving force of modern global environmental challenges, occurring at the end of the 20th Century, brought a new perspective to environmental and Earth system sciences. Lubchenco (1998) called directly for the integration of the human dimensions of global environmental changes with the physical-chemical-biological dimensions. In this context, Crutzen (2006) proposed the distinction of the Anthropocene as a new geological epoch, where the human species becomes a force outcompeting natural processes. As one possible framework to assess human agency in the Anthropocene, Schellnhuber (1999) developed the notion of “Earth System” analysis for global environmental management in which the human force has been conceptualized as a “global subject”. The global subject is a real but abstract force that represents the collective action of humanity as a self-conscious force that has conquered the planet. The global subject manifests itself, for instance, by adopting international protocols for climate protection.

The conceptualization of the human species as the global subject has been applied in Integrated Assessment Models (IAMs). IAMs refer to tools assessing strategies to address climate change and they aim to describe the complex relations between environmental, social and economic factors that determine future climate change and the effects of climate policy (van Vuuren et al., 2011). IAMs have been valuable means to set out potential pathways to mitigate climate change and, importantly, have been used in the IPCC's assessments of climate change mitigation (Clarke et al., 2014). However, the development of Integrated Assessment Models (IAMs) coincides in time with the supremacy of the rational choice paradigm. Rational choice theory emphasizes the voluntary nature of human action and the influence of such actions on decisions, assuming human beings act on the basis of rational calculations of benefits and costs (Burns, 1994). According to this paradigm, rationality is a feature of individual actors and the world can be explained in terms of interactions of atomic entities. Humans are rational beings motivated by self-interest and consciously evaluate alternative courses of action. Markets are seen as the mechanisms linking the micro and macro levels and allow the combination of the concrete actions of individuals, e.g. buyers and sellers (Jaeger et al., 2001). The rational choice paradigm is reflected in welfare maximization assumptions underpinning the development of computable general equilibrium (CGE) models that are widespread in IAMs. CGE models are computer-based simulations which use a system of equations that describe the whole world economy and their sectoral interactions. The analysis of scenarios in CGE models compares a business-as-usual equilibrium with the changes introduced by one or several policies and environmental shocks — e.g. a carbon tax or emissions trading scheme under several climate scenarios — which generate a new equilibrium (Babatunde et al., 2017). It is important to understand that the policy shock in such models is introduced externally; it does not evolve from the model and does not consider the dynamics behind the agency of different actors and groups of actors. In fact, human societies in CGE models are only reflected in aggregated population numbers by world region. The institutional settings within the human societies operate are given and cannot be endogenously changed. CGE models place a strong emphasis on the market as a solution to all kinds of problems including environmental and social issues (Scrieciu, 2007). Furthermore, state-of-the-art IAMs model aggregate datasets of sub-continental size. For instance, the IAM known as REMIND considers just 11 world regions, while the energy component of IMAGE considers only 26. The order of magnitude of the population of each of these regions is between 287 M and 680 M inhabitants (ADVANCE, 2017). Similarly, in the global land use allocation model MAgPIE, the food energy demand for ten types of food energy categories (cereals, rice, vegetable oils, pulses, roots and tubers, sugar, ruminant meat, non-ruminant meat, and milk) in ten world regions differentiated in the model is determined exogenously by population size and income growth, assuming that, for example, higher income is related to a higher demand for meat and milk (Popp et al., 2010). The impacts of changing lifestyles and the implications of demand-side solutions can be explored only manually by varying the underlying assumptions.

In context of the definition of human agency used above, IAMs reflect an agency of a rational consumer who decides on the choice of an optimal action having access to perfect information about the alternatives. By analyzing energy, land use, and their implications on global emissions (e.g. van Vuuren et al., 2012; Hibbard et al., 2010) IAMs can compute an economic setup to maximize welfare functions. Nevertheless, the welfare functions do not cover the diversity of human preferences. Complex distinctions of qualitative aspects, such as networks or influencers that can drive these processes, do not exist.

This drawback has been noted by the IAM community and attempts have been made to integrate human agency related behavior towards the political economy, social behavioral and interaction patterns (Riahi et al., 2017), or regimes of effort sharing (van den Berg et al., 2019) have been made. Some models also consider inequality and a diversity of consumption patterns (Hasegawa et al., 2015; McCollum et al. 2018). However, these approaches are still driven by exogenous quantifications and are unable to sufficiently inspect dynamics of human agency. Although IAMs are able to design pathways combining multiple strategies to achieve the 1.5 °C target of the Paris Agreement, which include human agency related actions such as lifestyle changes (van Vuuren et al., 2018), many questions remain. For example, how can human agency be triggered to achieve the lifestyle changes, at an individual level, necessary to achieve the 1.5 °C target? Also, how can the necessary institutional dynamics be brought into play? So far, these aspects are rarely considered in IAMs.

Novel and promising modelling approaches to incorporate human agency are being developed in complex network science (Borgatti et al., 2009) and social-ecological system modelling (Pérez et al., 2016). Complex networks usually consist of a set of nodes representing individual agents or representative aggregations thereof (such as business parties, geographical regions or countries) which are connected by different types of linkages, such as business relations, diplomatic ties, or even acquaintance and friendship (Newman, 2018). This type of

Reproduced from: I. M. Otto et al., Human agency in the Anthropocene, Ecol. Econ., vol. 167, 2020, doi: 10.1016/j.ecolecon.2019.106463, with the permission of Elsevier

framework has been developed in the past, and applied successfully to describe heterogeneous datasets from the social sciences, and to establish conceptual models for socio-economic and socio-ecological dynamics (Filatova et al., 2013). Nevertheless, most of such models are still based on theoretical assumptions with weak links to empirical data. A closer link with empirical data has so far only been achieved at case study level, focusing on particular local socio-environmental phenomena such as fishery or water management with agents representing local resource users or managers (e.g. Suwarno et al., 2018; Troost and Berger, 2015). The questions driving this work are: (i) how can similar models be conceptualized in order to represent the whole World-Earth system of human societies and their bio-physical environment (Donges et al., 2018) and (ii) how can they be linked with empirical data?

3. The concept of human agency in social sciences

Dellas et al. (2011) refer to agency in the governance of the Earth system as the capacity to act in the face of earth system transformation or to produce effects that ultimately shape natural processes. Agency in Earth system governance may be considered as contributing to problem solving, or alternatively it could include the negative consequences of the authority to act. Lister (2003) and Coulthard (2012), in their research on agency related to environmental and citizenship problems, distinguish two dimensions: (i) 'everyday agency' being the daily decision-making around how to make ends meet, and 'strategic agency' involving long-term planning and strategies; and (ii) 'personal agency' which reflects individual choices and 'political and citizenship agency' which is related to the capacity of people to affect the wider change (Lister 2003). Personal agency varies significantly across human individuals. However, there are powerful examples of social protests and movements demonstrating that even individually disempowered people can have a strong voice if they act collectively (Kashwan, 2016). In the context of natural resources and environmental management, there are empirical examples of self-organized local and regional communities and grassroots movements crafting new institutions that limit the control of national authorities (García-López, 2018; Dang, 2018). To give an example, civil society groups in Mexico managed to shape the REDD+ policies to protect the rights of agrarian communities (Kashwan, 2017a). In this context, Bandura (2006) proposes the differentiation of individual, proxy and collective agency (2006: 165). Individual agency refers to situations in which people bring their influence to bear through their own actions. This varies substantially from person to person with respect to individual freedom to act and the consequences of action. Individual agency is influenced by a whole set of socio-economic characteristics including gender, age, education, religion, social, economic and political capital. In many cultures, the individual agency of women is limited, for example, by inheritance law or by informal norms restricting their mobility or educational opportunities (Otto et al., 2017). However, individual agency also varies with an individual's ability to change the system rules. For example, very wealthy or influential people might find it easier to set new market trends or influence public decision-making processes than those with fewer resources (Otto et al., 2019). Proxy, or socially mediated agency refers to situations in which individuals have no direct control over conditions that affect their lives, but they influence others who have the resources, knowledge, and means to act on their behalf to secure the outcome they desire. Collective agency refers to situations in which individuals pool their knowledge, skills, and resources, and act in concert to shape their future (Bandura 2006: 165). These dimensions of agency are visualized in Fig. 1.

The dominant view of human agency in Earth system science and integrated modelling approaches has so far focused on the left upper corner of Fig. 1, i.e. on the everyday agency of individual human agents. This would correspond to, for example, modelling the effects of food consumption on land use patterns (e.g. Popp et al., 2010). Interestingly, although opinion formation and election models are well advanced in game theory (e.g. Penn, 2009; Ding et al., 2010), they have not yet been applied to the formation of international environmental policy in IAMs. At the same time the recent so-called protest voting shows that a small fraction of voters can push public policy down a radically different pathway. Some studies link the protest voting and rising populism with increasing inequalities and the political and social exclusion of the poor and underprivileged (Becker et al., 2017). In some cases, radical policy changes might also be achieved by individual acts of civil disobedience and, in a destructive manner, by terrorist attacks. Civil disobedience represents the peaceful breaking of unjust or unethical laws and is a technique of resistance and protest whose purpose is to achieve social or political change by drawing attention to problems and influencing public opinion. Terrorism is defined as an act of violence for the purpose of intimidating or coercing a government or civilian population.

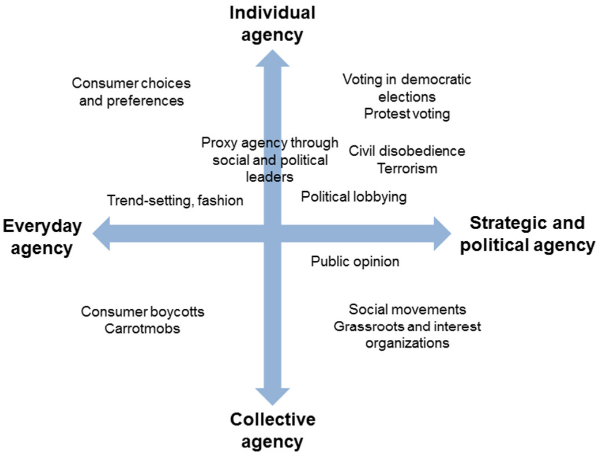


Fig. 1. Agency dimensions.
Adapted from Lister (2004) and Coulthard (2012) with empirical examples of social phenomena.

Furthermore, radical policy changes and social tipping points can emerge due to changes in the collective behavior and preferences. The term 'tipping point' "refers to a critical threshold at which a tiny perturbation can qualitatively alter the state or development of a system" (Lenton et al., 2008), hence the mere existence of tipping points implies that small perturbations created by parts of such a system can push the whole system into a different development trajectory. Examples of tipping-like phenomena in socio-economic systems include financial crises, but could also include the spread of new social values, pro-environmental behavior, social movements, and technological innovations (Steffen et al., 2018). To give an example, social movements and grassroots organizations played an important role in the German energy transition that was initiated in 2011 as a reaction to the nuclear disaster in Fukushima in Japan. It was, however, preceded by about 30 years of environmental activism (Hake et al., 2015). Finally, tipping-like phenomena can also be achieved by consumer boycotts and carrotmob movements. Consumer boycotts coupled with environmental NGO campaigns led, in Europe, to changes in the animal welfare regulations and the implementation of fair trade schemes (Belk et al., 2005). Carrotmobs refer to consumers collectively swarming a specific store to purchase its goods in order to reward corporate socially responsible behavior (Hoffmann and Hutter, 2012).

At the same time, cultural values and the ethical interpretation of behavior might vary in some respects across different countries and world regions and will lead to different manifestations of agency. Cultural values provide a strong filter of the actions perceived as good or responsible, as well as the consequences of violating norms (Belk et al., 2005). In the climate change context, some authors link the

Reproduced from: I. M. Otto et al., Human agency in the Anthropocene, Ecol. Econ., vol. 167, 2020, doi: 10.1016/j.ecolecon.2019.106463, with the permission of Elsevier

public acceptance of climate policy instruments to the belief and value systems in place, and the perceptions of the environment (Otto-Banaszak et al., 2011).

4. The manifestation of human agency: the layers of social structure

Biermann and Siebenhüner (2009) propose a distinction between actors and agents in Earth system governance. Actors are the individuals, organizations, and networks that participate in the decision-making processes. Agents are those actors who have the ability to prescribe behavior. The collective prescriptions and constraints on human behavior are usually referred to as the social structure (Granovetter, 1985; Dobres and Robb, 2000). The social structure is composed of the rule system that constitutes the "grammar" for social action that is used by the actors to structure and regulate their transactions with one another in defined situations or spheres of activity. The complex and multidimensional normative network is not given, but is a product of human action; "human agents continually form and reform social rule systems" (Burns and Flam, 1986: 26). The social rule system can also be framed as social institutions that are involved in political, economic, and social interactions (North, 1991). Similarly, Elinor Ostrom defines institutions as "the prescriptions that humans use to organize all forms of repetitive and structured interactions. Individuals interacting within rule-structured situations face choices regarding the actions and strategies they take, leading to consequences for themselves and for others" (Ostrom, 2005: 3). Social norms are shared understandings of actions and define which actions are obligatory, permitted, and forbidden (Crawford and Ostrom, 1995). Social order is only possible insofar as participants have common values and they share an understanding of their common interests and goals (King, 2009). Williamson (1998) proposes differentiating different informal institutions such as norms, beliefs and traditions, and formal institutions that comprise formal and written codes of conduct.

The process of shaping of the social rule system formation is not always fully conscious and intended. Lloyd (1988: 10) points out that a social structure is emerging from intended and unintended consequences of individual action and patterned mass behavior over time "Once such structures emerge, they feedback on the actions" (Sztompka, 1991: 49). For Giddens (1984) human action occurs as a continuous flow of conduct and he proposed turning the static notion of structure into the dynamic category of structuration to describe the human collective conduct. Human history is created by intentional activities but it is not an intended project; it persistently eludes efforts to bring it under conscious direction (Giddens, 1984: 27). As pointed out by Sztompka (1994), Giddens, embodies human agency in the everyday conduct of common people who are often distant from reformist intentions but are still involved in shaping and reshaping human societies. This process of the formation of social structure takes place over time; the system which individuals follow today have been produced and developed over a long period. "Through their transactions social groups and communities maintain and extend rule systems into the future" (Burns and Flam 1987: 29).

Another element of the social structure that is identified by several authors corresponds to the network of human relationships that, just like the shapes in geometry, can take different forms and configurations (Simmel, 1971). The network of relationships among the social agents is also referred to as governance structures, or sometimes as organizations. North (1990: 73) defines organizations as "purposive entities designed by their creators to maximize wealth, income, or other objectives defined by the opportunities afforded by the institutional structure of the society." Williamson (1998), focusing on the types of contracts, distinguishes three basic types of governance structures: markets, firms, and hybrids. In markets, transaction partners are autonomous; in firms, partners are inter-dependent and integrated into an internal organization. Hybrids are intermediate forms in which contract partners are bilaterally dependent but to a large degree maintain autonomy (Williamson 1996: 95–98). Studying communication networks and social group structures allows us to distinguish more social network relationship patterns (Sztompka, 2002: 138).

Finally, the social structure is also shaped and influenced by large material objects such as infrastructure and other technological and industrial structures, that some authors call the technosphere (Spaargaren, 1997: 78). Herrmann-Pillath (2018) defines the technosphere as the encompassing aggregate of all artificial objects in opposition to the natural world, and more specifically, establishes the systemic separateness of the technosphere relative to the biosphere. Just as social norms impose on one hand certain constrains on human behavior, however, on the other hand, structure the human interactions and also provide certain opportunities, the technosphere can be viewed as a humanly designed constructs that provide certain opportunities as well as they limit certain choices of individuals operating at different geographical and time scales (Donges et al., 2017a).

The system is fully interconnected, and the social structure layers are interrelated. The slow changing layers of social structure impose constraints on the layers that change more quickly. The faster changing layers of social structure, however, are also able to change the slow slayers through feedback mechanisms (c.f. Williamson, 2000). Human agency is manifested through the maintenance, reproduction and modifications in the social structure layers (Burns, 1994). Interestingly, infrastructure objects in the technosphere layer show a similar order of change as the informal and formal institutions, and thus might constrain the social change in the faster changing levels. Thus artefacts become co-carriers of agency (Herrmann-Pillath, 2018). Nevertheless, sharp brakes from the established procedures rarely happen. Such defining moments are an exception to the rule and usually emerge from massive discontents such as civil wars, revolutions, or financial crises (Williamson, 1998). Institutions can also lock the society into a path-dependence (Beddoe et al., 2009). The capacity to undergo a radical restructuring, however, is a unique feature distinguishing social systems from organic or mechanical ones. Restructuring the social structure is a product of human agency and is grounded in the interaction between structures and human actions that produces change in a system's given form, structure or state (Archer, 1988: xxii). However, the transition of institutions is frequently driven by crises (Beddoe et al., 2009).

Burns (1994: 215-216) introduces the notion of 'windows of opportunity' that are very relevant for analyzing social transformations. Interactive situations lacking social equilibria, which typically occur after catastrophes and other shocks, usually give rise to uncertainty, unpredictability, and confusion, and motivate actors to try, individually or collectively, to restructure the situation. In such restructuring activities, actors typically engage in reflective processes and make "choices about choice" and participate in meta-games (Burns 1994: 208). The actors may structure and restructure their preferences, outcomes, and outcome structures, and occasionally also the entire decision and game systems in which they participate. Through such structuring activity, human agents also create, maintain and change institutions and collective or organized agents such as movements, the state, market and bureaucratic organizations (Burns and Dietz, 1992; Burns, 1994: 215–216).

Transformations are the moments in history when the meta choices - "choices about choices" are made. The outcomes of such choices and the new type of system depend largely on the agents that get involved in the collective process of designing the new system. This process could be exclusive and incorporate only a narrow group of decision-makers as frequently happens in "quiet" transitions to authoritarian regimes. Alternatively, they can be more open and include representatives of various social groups, as happened in the political and economic transformation in Eastern Europe. Taking this example, Burns (1994) proposes that transformations are a co-evolutionary process sometimes driven by contradicting actors' interests. Transformations might entail shifts in core societal organizing principles and systems of rules. As a

Reproduced from: I. M. Otto et al., Human agency in the Anthropocene, Ecol. Econ., vol. 167, 2020, doi: 10.1016/j.ecolecon.2019.106463, with the permission of Elsevier

Table 1
The layers of social structure, the dominant type of agency and the order of change.
(Following Williamson, 1998).

Structure layer	Sub-components	The dominant type of human agency	The order of change
Institutional	Informal rules: norms, religion, tradition, customs	Collective and strategic	30 to over 100 years
	Formal rules: constitutions, written codes of conduct, judiciary, property rights	Collective and citizenship	10 to 50 years
Organizational	Governance structures	Proxy and strategic	5 to 10 years
	Organizations	Proxy, strategic	5 to 10 years
	Networks	Proxy, individual, everyday	Continuous
Technosphere	Infrastructure	Proxy, strategic	10–50 years
	Technology	Proxy, individual and everyday	Continuous

result, agents with vested interests may struggle to maintain established systems or to limit the changes within them. Other agents act openly or covertly to modify or transform the system. Table 1 summarizes the above discussion and tries to link the social structure layers to the dominant type of human agency that can to be used to transform them.

Even in periods of radical change, however, the actors never start from scratch. They cannot choose a completely new system and they always depart from the ongoing social order in which they are embedded. The future evolves from practical activities, experiments, learning, conflict and struggle (Burns, 1994: 216). A similar point of view is presented by evolutionary institutional economists, in which transformations are seen not as a simple replacement of old institutions by new ones, but as a recombination and reworking of old and new elements and groups of actors (e.g. Stark, 1996; Bromley, 2000).

5. Distribution of human agency: differentiating socio-metabolic agent classes

Following the rational choice paradigm could lead us to a conclusion that the society is a sum of individuals (Burns, 1994) and that any forms of agency should be equally distributed among the individuals in the society. Such an approach is typical for integrated assessment models in which human systems are usually separated into population and economic sectors. The parameters that describe population are usually mainly population number, and economic production determines the use of resources and pollution emissions in the model (e.g. van Vuuren et al., 2012).

It is, however, enough to observe the world to know that such assumptions are very simplistic. People's resource use and pollution emissions differ according to income, place of abode, type of occupation, and possessions. Moreover, their goals and interests, and the likelihood of them being fulfilled also differ. There are powerful individuals and groups in society who successfully strive for their interests, and there are individuals and groups who, despite struggling, never achieve their objectives. There are also masses of individuals who just strive to make ends meet. The questions are what types of agents or organizations can be incorporated in the models and what sort of agency do they have? Is there a need for a new social class theory taking access to energy and related carbon emissions as the base of social stratification?

Most social differentiation theories follow either the Marxist distinction between physical and capital endowments or the Weberian approach which differentiates classes through inequalities in ownership and income (Kozyr-Kowalski, 1992: 53). Some class theorists also highlight the development stages and inequalities across different countries and world-regions (Offe, 1992: 122). One more dimension that has not been discussed so far by social differentiation theories is the socio-metabolic profile of social classes, which constitutes the common ground for social and natural sciences. Social metabolism refers to the material flows in human societies and the way societies organize their exchanges of energy and materials with the environment (Fischer-Kowalski, 1997; Martinez-Alier, 2009). Social classes can be differentiated based on their metabolic profiles (Martinez-Alier, 2009). The use of energy by human beings can be divided into two main categories. The first one refers to the endosomatic use of energy as food, and the second one refers to the exosomatic use of energy as fuel for cooking and heating, and as power for the artefacts and machines produced by human society. Thus one person a day must eat the equivalent of 1500 to 2500 kcal to sustain their life functions, which is equivalent to about 10 MJ (megajoules) of energy per day or 3.65 GJ per year (Martinez-Alier, 2009). This amount varies only slightly among human beings. A rich person physically cannot eat much more, and even poorer individuals need the equivalent energy in the form of food to survive. Dietary composition and the amount of waste produced, however, will differ across the social strata. Nevertheless, there are still people suffering from hunger, unable to meet their basic needs.

The exosomatic energy use varies to a greater degree. The poorest social groups, who have no permanent access to electricity in their homes, who obtain energy for cooking and heating from the combustion of biomass products, who use overcrowded buses and trains to travel, use in total about 10 GJ of energy per person per year (Martinez-Alier, 2009) and constitute the lowest, socio-metabolic underclass. A more detailed picture can be derived by comparing the carbon footprint of different socio-economic groups. Personal CO_2 emissions are released directly in fuel combustion processes in vehicles, airplanes, heating and cooking appliances, and indirectly through electricity use and consumption of products that generated emissions in the upstream production processes. The authors include CO_2 emissions from energy used directly in homes (for space heating, lighting, etc.), for personal transportation (including personal vehicles and passenger aviation), and from the energy embedded in the production of goods consumed. Kümmel (2011) proposes the term "energy slaves" to describe the exosomatic energy use from fossil fuels by modern human society. On average, the daily energy consumption of a human being is equivalent to the men power of 15 people. Inhabitants of the most energy intensive Western Societies (i.e. the U.S.) consume, per person, the equivalent of the work of 92 people every day.

The results from UK households show that CO_2 emissions are strongly income, but also location, dependent. The highest emissions can be generated by people living in suburbs, mostly in detached houses, and having two or more cars. Emissions of such households equated to about 26 CO_2 tonnes in 2004. This amount was 64% higher than the emissions of the group with lowest emissions of 16 CO_2, which comprised mostly of older and single person urban households as well as the unemployed living mostly in urban areas (Druckman and Jackson, 2009). UK household emissions can be compared with emissions from households located in less developed countries. For example, household emissions in Malaysia, as in the UK, are strongly dependent on income and location. However, Malaysian households with the lowest emissions were found in villages as well as in low-income urban squatter settlements. The urban squatter settlement households emitted on average 10.18 CO_2 tonnes. The village households emitted on average 9.58 CO_2 tonnes per year. Households with the highest CO_2 emissions were located in high cost housing areas and they were responsible on average for 20.14 CO_2 tonnes per year (Majid et al., 2014).

On the other end of the social ladder, there are super-rich hyper-

Reproduced from: I. M. Otto et al., Human agency in the Anthropocene, Ecol. Econ., vol. 167, 2020, doi: 10.1016/j.ecolecon.2019.106463, with the permission of Elsevier

Table 2
Socio-metabolic class differentiation.
(Based on: Oxfam, 2015; Otto et al., 2019).

	Percent of global population	Percent of life-style CO_2 emissions	The level of human agency
Socio-metabolic underclass	20%	2.5%	Extremely low
Socio-metabolic energy poor class	30%	7.5%	Low
Socio-metabolic lower class	30%	22%	Moderate level of collective agency
Socio-metabolic middle class	10%	19%	Moderate to high
Socio-metabolic upper class	9.5%	35.4%	Very high
Super-rich	0.54%	13.6%	Extremely high

mobile individuals with multiple spacious residences, and whose livestyles are characterized by conspicuous consumption patterns. They are less than 1% of global population and their consumption related greenhouse gas emissions could be over 170 times higher than the world's poorest 10% (Oxfam, 2015). They can be characterized by extremely high levels of all types of agency. The influence and roles of many super-rich in the world of politics, media, culture, business and industry are often inter-related. In contrast to the super-rich in preindustrial societies they have almost unlimited mobility, owning properties in different counties, with their homes being guarded and fortified. They have the ability to switch countries of residence, taking the advantage of ‘nondomiciled’ tax status, i.e. being the national of a certain country while not actually living there (Paris, 2013). Table 2 presents a first attempt to stratify the global population according to their socio-metabolic profiles that is based on disaggregated data on consumption related carbon emissions (Oxfam, 2015; Otto et al., 2019).

The proportions in Table 2 are striking. The top 10% of the global population is responsible for almost 50% of global consumption related greenhouse gas emissions. The wealthiest 0.54% of the human population is responsible for more lifestyle carbon emissions than the poorest 50% (Otto et al., 2019).

Energy use, as well as carbon dioxide emission, can also be used to analyze the socio-metabolic profile of economic sectors, companies and other organizations. From 1854 to 2010 12.5% of all industrial carbon pollution was produced by just five companies – Chevron, ExxonMobil, British Petroleum, Shell and Conoco Philipps (Union of Concerned Scientists, 2018). To give an example from a different sector – in 2015 Saint-Gobain, a French multinational building materials manufacturer emitted 9.5 million metric tonnes CO_2e (Carbon Disclosure Project, 2016: 22). For a comparison, emissions from industrial processes in France in 2013 equated to 17.6 million tonnes CO_2e (General Directorate for Sustainable Development, 2016: 25) (GTM, 2018).

The socio-metabolic profile of social classes, nations, and organizations can be directly linked with their agency in the Earth system. The global socio-metabolic underclass is obviously characterized by a very low degree of agency. There are rare exceptions of mass protests initiated by the poorest social groups that can collectively influence formal institutions and change their governance (Kashwan, 2017b). However, these people are mostly occupied with making ends meet and have low organizational capabilities. In contrast, the global socio-metabolic upper classes are those who are characterized by a high level of individual agency as well as having the organizational capabilities to actively exercise their agency. Due to their resource incentive life-style they also have the moral obligation to be the agents of a transformation in global sustainability.

6. Improving the representation of human agency in integrated assessment modelling

In this section we ask how the above conceptual discussion could be summarized into guidelines improving the operationalization of human agency in Earth system science and integrated assessment modelling. In order to incorporate the different aspects of human agency as discussed in the previous sections, there is a need to introduce agents with heterogeneous goals, opinions and preferences into the models. The agents should be able to form networks that represent their mutual interrelationships and interactions between them. These system interaction rules should ideally refer to the social structure layers differentiated in Table 1, forming a nested hierarchical embeddedness of each agent.

Conceptual models, that incorporate the above requirements have been successfully developed and studied in the recent past. Their core properties might thus form a proper basis for extending IAMs to include heterogeneous agency on the level of (representative) individuals. Such models have been utilized to study opinion, and the associated consensus-formation specifically under the assumption of heterogeneous agents. Most of these works are based on the voter model in which agents exchange discrete (sets of) opinions in order to reach some consensus on a given (possibly abstract) topic or problem (Clifford and Sudbury, 1973; Holley and Liggett, 1975). Acknowledging that in its standard version the voter model considers all agents to have identical agency, extensions have been based on social impact theory (Latane, 1981) that specifically include heterogeneous relationships between single actors or groups (Nowak et al., 1990). Such extended models generally account for proximities between agents in some abstract space of personal relationships which is commonly modeled by assigning agents unique values of persuasiveness and supportiveness, describing their agency with respect to influencing as well as supporting others. While being of generic nature such classes of models can be easily modified to account for various kinds of processes related to social behavior, such as social learning (Kohring, 1996) or leadership (Holyst et al., 2001), which are again directly related to the notions of (heterogeneous distributions of) human agency. Certain models include additional layers of complexity by also accounting for the heterogeneous distribution of different group sizes (Sznajd-Weron, 2005) and certain majorities within those groups (Galam, 2002) when determining criteria for consensus in opinion dynamics.

One particular model of general cultural dynamics that has attracted great interest in the social science community, and that should be highlighted here, is the so-called Axelrod model (Axelrod, 1997). In its core, it accounts for two commonly observed tendencies in large groups of individuals or aggregations thereof: social influence (i.e. agency) and homophily (a process that dynamically influences each individual's agency over time). The Axelrod-model not only specifically accounts for heterogeneity in the different agents but also (and to some degree unintuitively) allows emerging cultural diversity to be modeled in its convergent state. In general, such flexible approaches allow incorporating individual human agency in terms of the different ties an agent might have with others (Emirbayer and Goodwin, 1994; Granovetter, 1977). Additionally, each tie can be associated with different strengths, thus also incorporating heterogeneity in the human agency (Castellano et al., 2009). Network modelling approaches further allow us to explicitly resolve the associated social structure (as well as the temporal evolution thereof) through an evaluation of the overall topology of the network on the meso- or macroscale (Costa et al., 2007).

A necessary step in operationalizing human agency in IAMs includes differentiating global socio-metabolic agent classes with heterogeneous metabolic profiles linking them with the material and energy flows in

Reproduced from: I. M. Otto et al., Human agency in the Anthropocene, Ecol. Econ., vol. 167, 2020, doi: 10.1016/j.ecolecon.2019.106463, with the permission of Elsevier

the bio-physical environment as well as heterogeneous social profiles that specify their preferences, opinions, and positions in social networks. Such efforts could be linked to the emerging research on downscaling planetary boundaries (Häyhä et al., 2016) as well as the established research on differentiating social milieus (e.g. Bauer and Gaskell, 1999). Some authors also propose model co-development, together with citizens and citizen groups (Figueres et al., 2017). Some authors also recommend abandoning the search for one gold-standard model, and instead explore future pathways based on a multitude of different concepts and representations of people and human agency (Donges et al., 2017b). For example, Donges et al. (2018) propose a modelling framework allowing incorporation of large sets of different models and concepts, in a standardized form, in order to assess and compare different future trajectories.

7. Conclusions

The Anthropocene has emerged unintentionally as a side effect of the industrialization of human societies (Crutzen, 2006). There are only a few examples of the human ability to internally interact with planetary geological forces, with the Montreal Protocol being the most often referred to example (Velders et al., 2007). At the same time historical examples show that there are instances of rapid transitions in societies (Bunker and Alban, 1997). Achieving policy challenges as outlined in the Sustainable Development Goals require a certain degree of societal transformation. The concept of agency is central to implementing transformations needed to limit global warming and achieve the SDGs. Most of the IAMs that dominate the scientific assessments of global environmental changes do not include a representation of human societies that would have a capacity to undertake system transformations. At the same time, there is a relatively rich social science theory that can be used to improve the operationalization of human agency in integrated assessment modelling efforts.

In this paper we show that human agency can actively shape the World-Earth system (c.f. Donges et al., 2018) through interventions at different layers of social structure. Human agency, however, is not evenly distributed across all human individuals and social groups. We postulate a differentiation of socio-metabolic agent classes that could be integrated into integrated assessment modelling efforts. More socio-economic sub-national and sub-population group data is needed for this purpose (c.f. Otto et al., 2015). Social institutions for sustainable management of global, regional, and local ecosystems, however, do not generally evolve spontaneously, but have to be consciously designed and implemented by the resource users (Gatzweiler and Hagedorn, 2002; Kluvankova-Oravska et al., 2009). Each social transformation contains a disruptive component that implies a destruction of existing patterns of social interaction and institutional structures, and creation and emergence of new patterns and structures. Introducing more dimensions of human agency into IAMs, and co-creating scenarios and pathways for modelling exercises together with citizens and institutions, would help break the barriers that disconnect peoples' actuality and agency with models, a discourse which has been gaining weight among policy makers (Figures, 2016). This disconnection can be broken by co-developing with citizens and various resource users the elements of global human-environmental system models, and by considering the people behind the numbers and the possible ways of funneling their agency. We encourage the integrated modelling community to work more closely with social scientists as well as we encourage social scientists to explore the methods and concepts applied in natural sciences.

Acknowledgments

The authors are grateful to two anonymous reviewers for their feedback that helped them to improve the paper. I.M.O., J.F.D., and R.C. are grateful for financial support by the Earth League's EarthDoc programme. I.M.O. is supported by funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 821010. J.F.D. is grateful for financial support by the Stordalen Foundation via the Planetary Boundary Research Network (PB.net). M.W. is financially supported by the Leibniz Association (project DominoES). This research has been carried out within the COPAN – Co-evolutionary Pathways Research Group at PIK.

References

ADVANCE, 2017. The common integrated assessment model (IAM) documentation. http://iamcdocumentation.eu/index.php/IAMC_wiki.

Archer, M.S., 1988. Culture and Agency. Cambridge University Press, Cambridge.

Axelrod, Robert, 1997. The Complexity of Cooperation. Agent-based Models of Competition and Collaboration. Princeton University Press, Princenton.

Babatunde, Kazeem Alasinrin, Begum, Rawshan Ara, Said, Fathin Faizah, 2017. Application of computable general equilibrium (CGE) to climate change mitigation policy: a systematic review. Renew. Sust. Energ. Rev. 78 (October), 61–71. https://doi.org/10.1016/j.rser.2017.04.064.

Bandura, Albert, 2006. Toward a Psychology of Human Agency. Perspect. Psychol. Sci. 1 (2), 164–180.

Bauer, Martin W., Gaskell, George, 1999. Towards a paradigm for research on social representations. J. Theory Soc. Behav. 29 (2), 163–186. https://doi.org/10.1111/1468-5914.00096.

Becker, Sascha O., Fetzer, Thiemo, Novy, Dennis, 2017. Who voted for brexit? A comprehensive district-level analysis. Econ. Policy 32 (92), 601–650. https://doi.org/10.1093/epolic/eix012.

Beddoe, Rachael, Constanza, Robert, Farley, Joshua, Garza, Eric, Kent, Jennifer, et al., 2009. Overcoming systemic roadblocks to sustainability: the evolutionary redesign of worldviews, institutions, and technologies. PNAS 106 (8), 2483–2489.

Belk, Russell, Devinney, Timothy, Eckhardt, Giana, 2005. Consumer ethics across cultures. Consum. Mark. Cult. 8 (3), 275–289. https://doi.org/10.1080/10253860500160411.

Biermann, F., Siebenhüner, B., 2009. Managers of Global Change: The Influence of International Environmental Bureaucracies. MIT Press.

Biermann, F., Abbott, K., Andresen, S., Bäckstrand, K., Bernstein, S., Betsill, M.M., Bulkeley, H., et al., 2012. Navigating the Anthropocene: improving earth system governance. Science 335 (6074), 1306–1307. https://doi.org/10.1126/science.1217255.

Biermann, Frank, Bai, Xuemei, Bondre, Ninad, Broadgate, Wendy, Chen, Chen-Tung Arthur, Dube, Opha Pauline, Erisman, Jan Willem, et al., 2016. Down to earth: contextualizing the Anthropocene. Glob. Environ. Chang. 39 (July), 341–350. https://doi.org/10.1016/j.gloenvcha.2015.11.004.

Borgatti, Stephen P., Mehra, Ajay, Brass, Daniel J., Labianca, Giuseppe, 2009. Network analysis in the social sciences. Science 323 (5916), 892–895. https://doi.org/10.1126/science.1165821.

Bromley, Daniel, 2000. Most Difficult Passage: The Economic Transition in Central and Eastern Europe and the Former Soviet Union. (In . Berlin).

Bunker, B.B., Alban, B.T., 1997. Large Group Interventions: Engaging the Whole System for Rapid Change. Jossey-Bass Publications, San Francisco.

Burns, T.R., Flam, H., 1987. The Shaping of Social Organization. Sage Publications, London.

Burns, Tom R., 1994. Two conceptions of human agency: rational choice theory and the social theory of action. In: Sztompka, Piotr (Ed.), Agency and Structure. Reorienting Social Theory. Gordon and Breach Science Publishers, Yverdon, Camberwell, Paris.

Burns, Tom R., Dietz, Thomas, 1992. Cultural evolution: social rule systems, selection and human agency, cultural evolution: social rule systems, selection and human agency. Int. Sociol. 7 (3), 259–283. https://doi.org/10.1177/026858092007003001.

Burns, Tom R., Flam, Helena, 1986. The Shaping of Social Organization: Social Rule System Theory and its Applications. Sage, London.

Carbon Disclosure Project, 2016. Embedding a carbon price into business strategy. . https://b8f65cb373b1b7b15feb-c70d8ead6ced550b4d987d7c03fcdd1d.ssl.cf3.rackcdn.com/cms/reports/documents/000/001/132/original/CDP_Carbon_Price_2016_Report.pdf?1474269757.

Castellano, Claudio, Fortunato, Santo, Loreto, Vittorio, 2009. Statistical physics of social dynamics. Rev. Mod. Phys. 81 (2), 591–646. https://doi.org/10.1103/RevModPhys.81.591.

Clarke, Leon, Akimoto, K., Babiker, M., et al., 2014. Assessing transformation pathways. Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, USA. https://www.ipcc.ch/pdf/assessment-report/ar5/wg3/ipcc_wg3_ar5_chapter6.pdf.

Clifford, Peter, Sudbury, Aidan, 1973. A model for spatial conflict. Biometrika 60 (3), 581–588. https://doi.org/10.1093/biomet/60.3.581.

Coleman, James S., 1990. Commentary: social institutions and social theory. Am. Sociol. Rev. 55 (3), 333–339. https://doi.org/10.2307/2095759.

Costa, L. da F., Rodrigues, F.A., Villas Boas, P.R., 2007. Characterization of complex networks: a survey of measurements. Adv. Phys. 59 (1), 167–242. https://doi.org/10.1080/00018730601170527.

Crawford, Sue E.S., Ostrom, Elinor, 1995. A grammar of institutions. Am. Polit. Sci. Rev. 89 (3), 582–600. https://doi.org/10.2307/2082975.

Crutzen, Paul J., 2006. The 'Anthropocene'. Earth System Science in the Anthropocene. Springer, Berlin, Heidelberg, pp. 13–18. https://doi.org/10.1007/3-540-26590-2_3.

Coulthard, S., 2012. Can we be both resilient and well, and what choices do people have?

Reproduced from: I. M. Otto et al., Human agency in the Anthropocene, Ecol. Econ., vol. 167, 2020, doi: 10.1016/j.ecolecon.2019.106463, with the permission of Elsevier

Incorporating agency into the resilience debate from a fisheries perspective. Ecol. Soc. 17 (1), 4.

Dang, Wenqi, 2018. How culture shapes environmental public participation: case studies of China, the Netherlands, and Italy. J. Chin. Gov. 0 (0), 1–23. https://doi.org/10.1080/23812346.2018.1443758.

Dellas, Eleni, Pattberg, Philipp, Betsill, Michele, 2011. Agency in earth system governance: refining a research agenda. Int. Environ. Agreements 11 (1), 85–98. https://doi.org/10.1007/s10784-011-9147-9.

Ding, Fei, Liu, Yun, Shen, Bo, Si, Xia-Meng, 2010. An evolutionary game theory model of binary opinion formation. Physica A 389 (8), 1745–1752. https://doi.org/10.1016/j.physa.2009.12.028.

Dobres, Marcia-Anne, Robb, John E., 2000. Ageny in Archaeology. Routledge, London and New York.

Donges Jonathan, F., Heitzig, Jobst, Barfuss, Wolfram, et al., 2018. Earth system modelling with complex dynamic human societies: the Copan:CORE World-Earth modeling framework. Earth Syst. Dyn. Disc. https://doi.org/10.5194/esd-2017-126.

Donges, J.F., Lucht, W., Müller-Hansen, F., Steffen, W., 2017a. The Technosphere in Earth system analysis: a coevolutionary perspective. Anthropocene Rev. 4 (1), 23–33. https://doi.org/10.1177/2053019616676608.

Donges, J.F., Winkelmann, R., Lucht, W., Cornell, S.E., Dyke, J.G., Rockström, J., Heitzig, J., Schellnhuber, H.J., 2017b. Closing the loop: reconnecting human dynamics to Earth system science. Anthropocene Rev. 4 (2), 151–157. https://doi.org/10.1177/2053019617725537.

Druckman, Angela, Jackson, Tim, 2009. The carbon footprint of UK households 1990-2004: a socio-economically disaggregated, quasi-mulit-regional input-output model. Ecol. Econ. 68, 2066–2077.

Edenhofer, O., Pichs-Madruga, R., Sokona, Y., 2014. Mitigation of Climate Change. Contribution of Work-ing Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, USA.

Emirbayer, Mustafa, Goodwin, Jeff, 1994. Network analysis, culture, and the problem of agency. Am. J. Sociol. 99 (6), 1411–1454. https://doi.org/10.1086/230450.

Figueres, Christiana, Schellnhuber, Hans Joachim, Whiteman, Gail, Rockström, Johan, Hobley, Anthony, Rahmstorf, Stefan, 2017. Three years to safeguard our climate. Nature 546 (7660), 593–595.

Figures, Christina, 2016. Plenary talk by Ch. Figueres, executive secretary of the United Nations Fremework convention on climate change. Plenary Talk at the Adaptation Futures Conference, Rotterdam, 10–13 May 2016, Amsterdam. https://www.youtube.com/watch?v=LOvDeIpVM8w.

Filatova, Tatiana, Verburg, Peter H., Parker, Dawn Cassandra, Stannard, Carol Ann, 2013. Spatial agent-based models for socio-ecological systems: challenges and prospects. *Environmental Modelling & Software*, Thematic Issue on Spatial Agent-Based Models for Socio-Ecological Systems, vol. 45, 1–7. https://doi.org/10.1016/j.envsoft.2013.03.017. (July).

Fischer-Kowalski, M., 1997. Society's metabolism: on the childhood and adolescence of a rising conceptual star. The International Handbook of Environmental Sociology. Edward Elgar Publishing, Cheltenham, UK.

Fuchs, Stephan, 2001. Beyond agency. Sociol. Theory 19 (1), 24–40. https://doi.org/10.1111/0735-2751.00126.

Galam, S., 2002. Minority opinion spreading in random geometry. Eur. Phys. J. B 25 (4), 403–406. https://doi.org/10.1140/epjb/e20020045.

García-López, Gustavo A., 2018. Rethinking elite persistence in neoliberalism: foresters and techno-bureaucratic logics in Mexico's community forestry. World Dev. https://doi.org/10.1016/j.worlddev.2018.03.018. March.

Gatzweiler, F.w., Hagedorn, K., 2002. The evolution of institutions in transition. Int. J. Agric. Resour. Gov. Ecol. 2 (1), 37–58. https://doi.org/10.1504/IJARGE.2002.000021.

General Directorate for Sustainable Development - SOeS, 2016. Key Figures on Climate France and Worldwide, 2016 edition. . http://www.statistiques.developpement-durable.gouv.fr/fileadmin/documents/Produits_editoriaux/Publications/Reperes/2015/highlights-key-figures-climate-2016-edition.pdf.

Giddens, Anthony, 1984. The Constitution of Society. Outline of the Theory of Structuration. University of California Press, Berkley and Los Angeles.

Granovetter, Mark, 1985. Economic action and social structure: the problem of embeddedness. Am. J. Sociol. 91 (3), 481–510. https://doi.org/10.1086/228311.

Granovetter, Mark S., 1977. The strength of weak ties11this paper originated in discussions with Harrison White, to whom I am indebted for many suggestions and ideas. Earlier drafts were read by Ivan Chase, James Davis, William Michelson, Nancy Lee, Peter Rossi, Charles Tilly, and an anonymous referee; their criticisms resulted in significant improvements. In: Leinhardt, Samuel (Ed.), Social Networks. Academic Press, pp. 347–367. https://doi.org/10.1016/B978-0-12-442450-0.50025-0.

GTM, 2018. Analysis: CO2 emissions at the World's largest companies are rising. May 2, 2018. https://www.greentechmedia.com/articles/read/analysis-firms-with-leading-climate-strategies-have-better-stock-performanc#gs.Y6dytuE.

Hake, Jürgen-Friedrich, Fischer, Wolfgang, Venghaus, Sandra, Weckenbrock, Christoph, 2015. The German Energiewende – history and status quo. *Energy*, Sustainable Development of Energy, Water and Environment Systems, vol. 92, 532–546. https://doi.org/10.1016/j.energy.2015.04.027. (December).

Hasegawa, Tomoko, Fujimori, Shinichiro, Takahashi, Kiyoshi, Masui, Toshihiko, 2015. Scenarios for the risk of hunger in the twenty first century using shared socio-economic pathways. Environ. Res. Lett. 10 (1).

Häyhä, Tiina, Lucas, Paul L., van Vuuren, Detlef P., Cornell, Sarah E., Hoff, Holger, 2016. From planetary boundaries to national fair shares of the global safe operating space — how can the scales be bridged? Glob. Environ. Chang. 40 (September), 60–72. https://doi.org/10.1016/j.gloenvcha.2016.06.008.

Herrmann-Pillath, Carsten, 2018. The case for a new discipline: technosphere science. Ecol. Econ. 149 (July), 212–225. https://doi.org/10.1016/j.ecolecon.2018.03.024.

Hibbard, Kathy, Janetos, Anthony, van Vuuren, Detlef P., Pongratz, Julia, Rose, Steven K., Betts, Richard, Herold, Martin, Feddema, Johannes J., 2010. Research priorities in land use and land-cover change for the earth system and integrated assessment modelling. Int. J. Climatol. 30 (13), 2118–2128. https://doi.org/10.1002/joc.2150.

Hoffmann, Stefan, Hutter, Katharina, 2012. Carrotmob as a new form of ethical consumption. The nature of the concept and avenues for future research. J. Consum. Policy 35 (2), 215–236. https://doi.org/10.1007/s10603-011-9185-2.

Holley, Richard A., Liggett, Thomas M., 1975. Ergodic theorems for weakly interacting infinite systems and the voter model. Ann. Probab. 3 (4), 643–663.

Holyst, Janusz A., Kacperski, Krzysztof, Schweitzer, Frank, 2001. Social impact models of opinion dynamics. Annual Reviews of Computational Physics IX, vol. 9, 253–273. Annual Reviews of Computational Physics, Volume 9. WORLD SCIENTIFIC. https://doi.org/10.1142/9789812811578_0005.

Jaeger, Carlo C., Renn, Ortwin, Rosa, Eugene A., Webler, Thomas, 2001. Risk, Uncertainty, and Rational Action. Earthscan Publications Ltd, London and Sterling.

Kashwan, Prakash, 2016. Inequality, democracy, and the environment: a cross-national analysis. Ecol. Econ. (131), 139–151.

Kashwan, Prakash, 2017a. Democracy in the woods. Envrionmental Conservation and Social Justice in India, Tanzania, and Mexico. Oxford University Press, New York.

Kashwan, Prakash, 2017b. Inequality, democracy, and the environment: a cross-national analysis. Ecol. Econ. 131, 139–151.

Kashwan, Prakash, MacLean, Lauren M., García-López, Gustavo A., 2018. Rethinking power and institutions in the shadows of neoliberalism: (an introduction to a special issue of world development). World Dev. https://doi.org/10.1016/j.worlddev.2018.05.026. May.

King, Anthony, 2009. Overcoming structure and agency. Talcott Parsons, Ludwig Wittgenstein and the theory of social action. J. Class. Sociol. 9 (2), 260–288.

Kluvankova-Oravska, Tatiana, Chobotova, Veronika, Slavikova, Lenka, Trifunovova, Sonja, 2009. From government to governance for biodiversity: the perspective of central and eastern European transition countries - 0f3175391c2efbd10d000000.Pdf. Environ. Policy Gov. 19, 186–196.

Kohring, G.A., 1996. Ising models of social impact: the role of cumulative advantage. J. Phys. I 6 (2), 301–308. https://doi.org/10.1051/jp1:1996150.

Kozyr-Kowalski, Stanislaw, 1992. Economic ownership and partial and combined classes of society. An attempt at a postive critique of post-Marxian Marxism. On Social Differentation. A Contribution to the Critique of Marxis Ideology. Adam Mickiewicz University Press, Poznan.

Kümmel, Reiner, 2011. The Second Law of Economics: Energy, Entropy, and the Origins of Wealth. Springer Science & Business Media.

McCollum, David L., Wilson, Charlie, Bevione, Michela, Carrara, Samuel, Edelenbosch, Oreane Y., Emmerling, Johannes, Guivarch, Céline, et al., 2018. Interaction of consumer preferences and climate policies in the global transition to low-carbon vehicles. Nat. Energy 1https://doi.org/10.1038/s41560-018-0195-z. July.

Nowak, A., Szamrej, J., Latané, B., 1990. From private attitude to public opinion: a dynamic theory of social impact. Psychol. Rev. 97 (3), 362.

Latane, Bibb, 1981. The psychology of social impact. Am. Psychol. 36 (4), 343–356.

Lenton, T.M., Held, H., Kriegler, E., Hall, J.W., Lucht, W., Rahmstorf, S., Schellnhuber, H.J., 2008. Tipping elements in the earth's climate system. Proc. Natl. Acad. Sci. 105 (6), 1786–1793. https://doi.org/10.1073/pnas.0705414105.

Lister, Ruth, 2003. What is citizenship? In: Lister, Ruth, Campling, Jo (Eds.), Citizenship: Feminist Perspectives. Macmillan Education UK, London, pp. 13–42. https://doi.org/10.1007/978-0-230-80253-7_2.

Lloyd, C., 1988. Explanation in Social History. Blackwell, Oxford.

Lubchenco, Jane, 1998. Entering the century of the environment: a new social contract for science. Science 279 (5350), 491–497. https://doi.org/10.1126/science.279.5350.491.

Majid, M.R., Moeinzadeh, S.N., Tifwa, H.Y., 2014. Income-carbon footprint relationship for urban and rural households of Iskandar Malaysia. IOP Conf. Ser. 18 (012164), 1–5.

Martinez-Alier, Joan, 2009. Social metabolism, ecological distribution conflicts, and languages of valuation. Capital. Nat. Social. 20 (1), 58–87.

Moss, Richard H., Edmonds, Jae A., Hibbard, Kathy A., Manning, Martin R., Rose, Steven K., et al., 2010. The next generation of scenarios for climate change research and assessment. Nature 463, 747–756. https://doi.org/10.1038/nature08823.

Newman, Mark, 2018. Networks. Oxford University Press.

North, Douglass C., 1990. Institutions, Institutional Change and Economic Performance. Cambridge University Press, Cambridge.

North, Douglass C., 1991. Institutions. J. Econ. Perspect. 5 (1), 97–112. https://doi.org/10.1257/jep.5.1.97.

Offe, Clauss, 1992. The welfare state, new class differentiation and diminution of social conflict. On Social Differentation. A Contribution to the Critique of Marxis Ideology. Adam Mickiewicz University Press, Poznan, pp. 97–130.

Ostrom, Elinor, 2005. Understanding Institutional Diversity. Princeton University Press, Princenton.

Otto, Ilona M., Biewald, Anne, Coumou, Dim, Feulner, Georg, Köhler, Claudia, Nocke, Thomas, Blok, Anders, et al., 2015. Socio-economic data for global environmental change research. Nat. Clim. Chang. 5, 503–506.

Otto, Ilona M., Reckien, Diana, Reyer, Christopher P.O., Marcus, Rachel, Masson, Virginie Le, Jones, Lindsey, Norton, Andrew, Serdeczny, Olivia, 2017. Social vulnerability to climate change: a review of concepts and evidence. Reg. Environ. Chang. https://doi.org/10.1007/s10113-017-1105-9. February.

Otto, Ilona M., Kim, Kyoung Mi, Dubrovsky, Nika, Lucht, Wolfgang, 2019. Shift the focus from the super-poor to the super-rich. Nat. Clim. Chang. 9 (2), 82–84. https://doi.org/10.1038/s41558-019-0402-3.

Otto-Banaszak, Ilona, Matczak, Piotr, Wesseler, Justus, Wechsung, Frank, 2011. Different

Reproduced from: I. M. Otto et al., Human agency in the Anthropocene, Ecol. Econ., vol. 167, 2020, doi: 10.1016/j.ecolecon.2019.106463, with the permission of Elsevier

perceptions of adaptation to climate change: a mental model approach applied to the evidence from expert interviews. Reg. Environ. Chang. 11 (2), 217–228. https://doi.org/10.1007/s10113-010-0144-2.

Oxfam, 2015. Extreme carbon inequality. Oxfam Media Briefing. https://www.oxfam.org/en/research/extreme-carbon-inequality.

Paris, Chris, 2013. The homes of the super-rich: multiple residences, hyper-mobility and decoupling of prime residential housing in global cities. Geographes of the Super-rich. Edward Elgar, Cheltenham, UK; Northhampton, MA, USA.

Pattberg, Philipp, Stripple, Johannes, 2008. Beyond the public and private divide: remapping transnational climate governance in the 21st century. Int. Environ. Agreements 8 (4), 367–388. https://doi.org/10.1007/s10784-008-9085-3.

Penn, Elizabeth Maggie, 2009. A model of farsighted voting. Am. J. Polit. Sci. 53 (1), 36–54. https://doi.org/10.1111/j.1540-5907.2008.00356.x.

Pérez, Irene, Janssen, Marco A., Anderies, John M., 2016. Food security in the face of climate change: adaptive capacity of small-scale social-ecological systems to environmental variability. Glob. Environ. Chang. 40 (September), 82–91. https://doi.org/10.1016/j.gloenvcha.2016.07.005.

Popp, Alexander, Lotze-Campen, Hermann, Bodirsky, Benjamin, 2010. Food consumption, diet shifts and associated non-CO2 greenhouse gases from agricultural production. Glob. Environ. Chang. 20 (3), 451–462. https://doi.org/10.1016/j.gloenvcha.2010.02.001.

Riahi, Keywan, van Vuuren, Detlef P., Kriegler, Elmar, Jae Edmonds, Brian C.O., Shinichiro Fujimori, Neill, Bauer, Nico, et al., 2017. The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: an overview. Glob. Environ. Chang. 42 (January), 153–168. https://doi.org/10.1016/j.gloenvcha.2016.05.009.

Schellnhuber, H.J., 1999. Earth system' analysis and the second Copernican revolution. Nature 402 (6761supp), C19–C23. https://doi.org/10.1038/35011515.

Scrieciu, S.Serban., 2007. The inherent dangers of using computable general equilibrium models as a single integrated modelling framework for sustainability impact assessment. A critical note on Böhringer and Löschel (2006). Ecol. Econ. 60 (4), 678–684. https://doi.org/10.1016/j.ecolecon.2006.09.012.

Simmel, Georg, 1971. On Individuality and Social Forms. Selected Writings. The University of Chicago Press, Chicago and London.

Spaargaren, Gert, 1997. The Ecological Modernization of Production and Consumption. Essays in Environmental Sociology. Universtiy of Wageningen, Wageningen.

Stark, David, 1996. Recombinant property in east European capitalism. Am. J. Sociol. 101 (4), 993–1027. https://doi.org/10.1086/230786.

Steffen, Will, Rockström, Johan, Richardson, Katherine, Lenton, Timothy M., Folke, Carl, Liverman, Diana, Summerhayes, Colin P., et al., 2018. Trajectories of the earth system in the Anthropocene. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.1810141115. August, 201810141.

Suwarno, Aritta, van Noordwijk, Meine, Weikard, Hans-Peter, Suyamto, Desi, 2018. Indonesia's forest conversion moratorium assessed with an agent-based model of land-use change and ecosystem services (LUCES). Mitig. Adapt. Strateg. Glob. Chang. 23 (2), 211–229. https://doi.org/10.1007/s11027-016-9721-0.

Sznajd-Weron, Katarzyna, 2005. Sznajd Model and Its Applications. ArXiv:Physics/0503239, March. http://arxiv.org/abs/physics/0503239.

Sztompka, Piotr, 1991. Society in Action. The Theory of Social Becoming. University of Chicago Press, Chicago.

Sztompka, Piotr, 1994. Evolving focus on human agency in contemporary social theory. Agency and Structure. Reorienting Social Theory. Gordon and Breach Science Publishers, Yverdon, Camberwell, Paris.

Sztompka, Piotr, 2002. Socjologia: Analiza Spoleczenstwa. Znak, Krakow.

Troost, Christian, Berger, Thomas, 2015. Dealing with uncertainty in agent-based simulation: farm-level modeling of adaptation to climate change in southwest Germany. Am. J. Agric. Econ. 97 (3), 833–854. https://doi.org/10.1093/ajae/aau076.

Union of Concerned Scientists, 2018. Largest producers of industrial carbon emissions. May 2, 2018. https://www.ucsusa.org/global-warming/science-and-impacts/science/largest-producers-industrial-carbon-emissions.html#.Wumo_n–mUl.

United Nations, 2015a. Resolution Adopted by the General Assemly on 25 September 2015. Transforming Our World: The 2030 Agenda for Sustainable Development. http://www.un.org/ga/search/view_doc.asp?symbol = A/RES/70/1&Lang = E.

United Nations, 2015b. UN Sustainable Development Goals. United Nations. http://www.un.org/sustainabledevelopment/sustainable-development-goals/#.

van den Berg, Nicole J., van Soest, Heleen L., Hof, Andries F., den Elzen, Michel G.J., van Vuuren, Detlef P., Chen, Wenying, Drouet, Laurent, et al., 2019. Implications of Various Effort-sharing Approaches for National Carbon Budgets and Emission Pathways. Climatic Change, February. https://doi.org/10.1007/s10584-019-02368-y.

van Vuuren, Detlef P., Lowe, Jason, Stehfest, Elke, Gohar, Laila, Hof, Andries F., Hope, Chris, Warren, Rachel, Meinshausen, Malte, Plattner, Gian-Kasper, 2011. How well do integrated assessment models simulate climate change? Clim. Chang. 104 (2), 255–285. https://doi.org/10.1007/s10584-009-9764-2.

van Vuuren, Detlef P., Bayer, Laura Batlle, Chuwah, Clifford, Ganzeveld, Laurens, Hazeleger, Wilco, van der Hurk, Bart, van Noije, Twan, O'Neil, Brian, Strengers, Bart J., 2012. A comprehensive view on climate change: coupling of earth system and integrated assessment models. Environ. Res. Lett. 7, 1–10.

van Vuuren, Detlef P., Stehfest, Elke, Gernaat, David E.H.J., van den Berg, Maarten, Bijl, David L., de Boer, Harmen Sytze, Daioglou, Vassilis, et al., 2018. Alternative pathways to the 1.5 °C target reduce the need for negative emission technologies. Nat. Clim. Chang. 8 (5), 391–397. https://doi.org/10.1038/s41558-018-0119-8.

Velders, Guus J.M., Andersen, Stephen O., Daniel, John S., Fahey, David W., McFarland, Mack, 2007. The importance of the Montreal protocol in protecting climate. Proc. Natl. Acad. Sci. 104 (12), 4814–4819. https://doi.org/10.1073/pnas.0610328104.

Williamson, Oliver E., 1996. In: The Mechanisms of Governance. Oxford University Press.

Williamson, Oliver E., 1998. Transaction costs economics: how it works; where it is headed. De Economist 146 (1), 23–58.

Williamson, Oliver E., 2000. The new institutional economics: taking stock, looking ahead. J. Econ. Lit. 38 (3), 595–613. https://doi.org/10.1257/jel.38.3.595.

Reproduced from: I. M. Otto et al., Human agency in the Anthropocene, Ecol. Econ., vol. 167, 2020, doi: 10.1016/j.ecolecon.2019.106463, with the permission of Elsevier

PERSPECTIVE

Trajectories of the Earth System in the Anthropocene

Will Steffen[a,b,1], Johan Rockström[a], Katherine Richardson[c], Timothy M. Lenton[d], Carl Folke[a,e], Diana Liverman[f], Colin P. Summerhayes[g], Anthony D. Barnosky[h], Sarah E. Cornell[a], Michel Crucifix[i,j], Jonathan F. Donges[a,k], Ingo Fetzer[a], Steven J. Lade[a,b], Marten Scheffer[l], Ricarda Winkelmann[k,m], and Hans Joachim Schellnhuber[a,k,m,1]

Edited by William C. Clark, Harvard University, Cambridge, MA, and approved July 6, 2018 (received for review June 19, 2018)

We explore the risk that self-reinforcing feedbacks could push the Earth System toward a planetary threshold that, if crossed, could prevent stabilization of the climate at intermediate temperature rises and cause continued warming on a "Hothouse Earth" pathway even as human emissions are reduced. Crossing the threshold would lead to a much higher global average temperature than any interglacial in the past 1.2 million years and to sea levels significantly higher than at any time in the Holocene. We examine the evidence that such a threshold might exist and where it might be. If the threshold is crossed, the resulting trajectory would likely cause serious disruptions to ecosystems, society, and economies. Collective human action is required to steer the Earth System away from a potential threshold and stabilize it in a habitable interglacial-like state. Such action entails stewardship of the entire Earth System—biosphere, climate, and societies—and could include decarbonization of the global economy, enhancement of biosphere carbon sinks, behavioral changes, technological innovations, new governance arrangements, and transformed social values.

Earth System trajectories | climate change | Anthropocene | biosphere feedbacks | tipping elements

The Anthropocene is a proposed new geological epoch (1) based on the observation that human impacts on essential planetary processes have become so profound (2) that they have driven the Earth out of the Holocene epoch in which agriculture, sedentary communities, and eventually, socially and technologically complex human societies developed. The formalization of the Anthropocene as a new geological epoch is being considered by the stratigraphic community (3), but regardless of the outcome of that process, it is becoming apparent that Anthropocene conditions transgress Holocene conditions in several respects (2). The knowledge that human activity now rivals geological forces in influencing the trajectory of the Earth System has important implications for both Earth System science and societal decision making. While recognizing that different societies around the world have contributed differently and unequally to pressures on the Earth System and will have varied capabilities to alter future trajectories (4), the sum total of human impacts on the system needs to be taken into account for analyzing future trajectories of the Earth System.

Here, we explore potential future trajectories of the Earth System by addressing the following questions.

> Is there a planetary threshold in the trajectory of the Earth System that, if crossed, could prevent stabilization in a range of intermediate temperature rises?
>
> Given our understanding of geophysical and biosphere feedbacks intrinsic to the Earth System, where might such a threshold be?

[a]Stockholm Resilience Centre, Stockholm University, 10691 Stockholm, Sweden; [b]Fenner School of Environment and Society, The Australian National University, Canberra, ACT 2601, Australia; [c]Center for Macroecology, Evolution, and Climate, University of Copenhagen, Natural History Museum of Denmark, 2100 Copenhagen, Denmark; [d]Earth System Science Group, College of Life and Environmental Sciences, University of Exeter, EX4 4QE Exeter, United Kingdom; [e]The Beijer Institute of Ecological Economics, The Royal Swedish Academy of Science, SE-10405 Stockholm, Sweden; [f]School of Geography and Development, The University of Arizona, Tucson, AZ 85721; [g]Scott Polar Research Institute, Cambridge University, CB2 1ER Cambridge, United Kingdom; [h]Jasper Ridge Biological Preserve, Stanford University, Stanford, CA 94305; [i]Earth and Life Institute, Université catholique de Louvain, 1348 Louvain-la-Neuve, Belgium; [j]Belgian National Fund of Scientific Research, 1000 Brussels, Belgium; [k]Research Domain Earth System Analysis, Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany; [l]Department of Environmental Sciences, Wageningen University & Research, 6700AA Wageningen, The Netherlands; and [m]Department of Physics and Astronomy, University of Potsdam, 14469 Potsdam, Germany

Author contributions: W.S., J.R., K.R., T.M.L., C.F., D.L., C.P.S., A.D.B., S.E.C., M.C., J.F.D., I.F., S.J.L., M.S., R.W., and H.J.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: will.steffen@anu.edu.au or john@pik-potsdam.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1810141115/-/DCSupplemental.

Published online August 6, 2018.

If a threshold is crossed, what are the implications, especially for the wellbeing of human societies?

What human actions could create a pathway that would steer the Earth System away from the potential threshold and toward the maintenance of interglacial-like conditions?

Addressing these questions requires a deep integration of knowledge from biogeophysical Earth System science with that from the social sciences and humanities on the development and functioning of human societies (5). Integrating the requisite knowledge can be difficult, especially in light of the formidable range of timescales involved. Increasingly, concepts from complex systems analysis provide a framework that unites the diverse fields of inquiry relevant to the Anthropocene (6). Earth System dynamics can be described, studied, and understood in terms of trajectories between alternate states separated by thresholds that are controlled by nonlinear processes, interactions, and feedbacks. Based on this framework, we argue that social and technological trends and decisions occurring over the next decade or two could significantly influence the trajectory of the Earth System for tens to hundreds of thousands of years and potentially lead to conditions that resemble planetary states that were last seen several millions of years ago, conditions that would be inhospitable to current human societies and to many other contemporary species.

Risk of a Hothouse Earth Pathway

Limit Cycles and Planetary Thresholds. The trajectory of the Earth System through the Late Quaternary, particularly the Holocene, provides the context for exploring the human-driven changes of the Anthropocene and the future trajectories of the system (*SI Appendix* has more detail). Fig. 1 shows a simplified representation of complex Earth System dynamics, where the physical climate system is subjected to the effects of slow changes in Earth's orbit and inclination. Over the Late Quaternary (past 1.2 million years), the system has remained bounded between glacial and interglacial extremes. Not every glacial–interglacial cycle of the past million years follows precisely the same trajectory (7), but the cycles follow the same overall pathway (a term that we use to refer to a family of broadly similar trajectories). The full glacial and interglacial states and the ca. 100,000-years oscillations between them in the Late Quaternary loosely constitute limit cycles (technically, the asymptotic dynamics of ice ages are best modeled as pullback attractors in a nonautonomous dynamical system). This limit cycle is shown in a schematic fashion in blue in Fig. 1, *Lower Left* using temperature and sea level as the axes. The Holocene is represented by the top of the limit cycle loop near the label A.

The current position of the Earth System in the Anthropocene is shown in Fig. 1, *Upper Right* by the small ball on the pathway that leads away from the glacial–interglacial limit cycle. In Fig. 2, a stability landscape, the current position of the Earth System is represented by the globe at the end of the solid arrow in the deepening Anthropocene basin of attraction.

The Anthropocene represents the beginning of a very rapid human-driven trajectory of the Earth System away from the glacial–interglacial limit cycle toward new, hotter climatic conditions and a profoundly different biosphere (2, 8, 9) (*SI Appendix*). The current position, at over 1 °C above a preindustrial baseline (10), is nearing the upper envelope of interglacial conditions over the past 1.2 million years (*SI Appendix*, Table S1). More importantly, the rapid trajectory of the climate system over the past half-century along with technological lock in and socioeconomic inertia in human systems commit the climate system to conditions beyond the envelope of past interglacial conditions. We, therefore, suggest that the Earth System may already have passed one "fork in the road" of potential pathways, a bifurcation (near A in Fig. 1) taking the Earth System out of the next glaciation cycle (11).

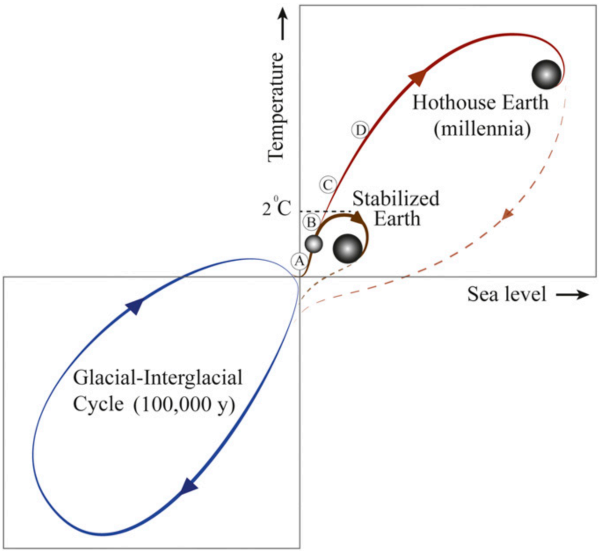


Fig. 1. A schematic illustration of possible future pathways of the climate against the background of the typical glacial–interglacial cycles (*Lower Left*). The interglacial state of the Earth System is at the top of the glacial–interglacial cycle, while the glacial state is at the bottom. Sea level follows temperature change relatively slowly through thermal expansion and the melting of glaciers and ice caps. The horizontal line in the middle of the figure represents the preindustrial temperature level, and the current position of the Earth System is shown by the small sphere on the red line close to the divergence between the Stabilized Earth and Hothouse Earth pathways. The proposed planetary threshold at ∼2 °C above the preindustrial level is also shown. The letters along the Stabilized Earth/Hothouse Earth pathways represent four time periods in Earth's recent past that may give insights into positions along these pathways (*SI Appendix*): A, Mid-Holocene; B, Eemian; C, Mid-Pliocene; and D, Mid-Miocene. Their positions on the pathway are approximate only. Their temperature ranges relative to preindustrial are given in *SI Appendix*, Table S1.

In the future, the Earth System could potentially follow many trajectories (12, 13), often represented by the large range of global temperature rises simulated by climate models (14). In most analyses, these trajectories are largely driven by the amount of greenhouse gases that human activities have already emitted and will continue to emit into the atmosphere over the rest of this century and beyond—with a presumed quasilinear relationship between cumulative carbon dioxide emissions and global temperature rise (14). However, here we suggest that biogeophysical feedback processes within the Earth System coupled with direct human degradation of the biosphere may play a more important role than normally assumed, limiting the range of potential future trajectories and potentially eliminating the possibility of the intermediate trajectories. We argue that there is a significant risk that these internal dynamics, especially strong nonlinearities in feedback processes, could become an important or perhaps, even dominant factor in steering the trajectory that the Earth System actually follows over coming centuries.

Reproduced from: W. Steffen et al., Trajectories of the Earth System in the Anthropocene, Proc. Natl. Acad. Sci. U. S. A., vol. 115, no. 33, pp. 8252–8259, 2018, doi: 10.1073/pnas.1810141115. Published under Creative Commons Attribution License 4.0 (CC BY NC ND).

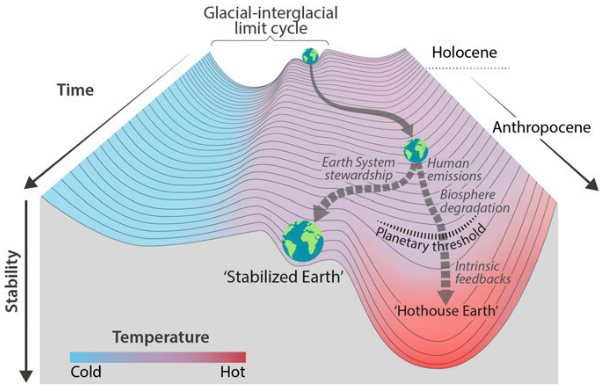


Fig. 2. **Stability landscape showing the pathway of the Earth System out of the Holocene and thus, out of the glacial–interglacial limit cycle to its present position in the hotter Anthropocene. The fork in the road in Fig. 1 is shown here as the two divergent pathways of the Earth System in the future (broken arrows). Currently, the Earth System is on a Hothouse Earth pathway driven by human emissions of greenhouse gases and biosphere degradation toward a planetary threshold at ∼2 °C (horizontal broken line at 2 °C in Fig. 1), beyond which the system follows an essentially irreversible pathway driven by intrinsic biogeophysical feedbacks. The other pathway leads to Stabilized Earth, a pathway of Earth System stewardship guided by human-created feedbacks to a quasistable, human-maintained basin of attraction. "Stability" (vertical axis) is defined here as the inverse of the potential energy of the system. Systems in a highly stable state (deep valley) have low potential energy, and considerable energy is required to move them out of this stable state. Systems in an unstable state (top of a hill) have high potential energy, and they require only a little additional energy to push them off the hill and down toward a valley of lower potential energy.**

This risk is represented in Figs. 1 and 2 by a planetary threshold (horizontal broken line in Fig. 1 on the Hothouse Earth pathway around 2 °C above preindustrial temperature). Beyond this threshold, intrinsic biogeophysical feedbacks in the Earth System (*Biogeophysical Feedbacks*) could become the dominant processes controlling the system's trajectory. Precisely where a potential planetary threshold might be is uncertain (15, 16). We suggest 2 °C because of the risk that a 2 °C warming could activate important tipping elements (12, 17), raising the temperature further to activate other tipping elements in a domino-like cascade that could take the Earth System to even higher temperatures (*Tipping Cascades*). Such cascades comprise, in essence, the dynamical process that leads to thresholds in complex systems (section 4.2 in ref. 18).

This analysis implies that, even if the Paris Accord target of a 1.5 °C to 2.0 °C rise in temperature is met, we cannot exclude the risk that a cascade of feedbacks could push the Earth System irreversibly onto a "Hothouse Earth" pathway. The challenge that humanity faces is to create a "Stabilized Earth" pathway that steers the Earth System away from its current trajectory toward the threshold beyond which is Hothouse Earth (Fig. 2). The human-created Stabilized Earth pathway leads to a basin of attraction that is not likely to exist in the Earth System's stability landscape without human stewardship to create and maintain it. Creating such a pathway and basin of attraction requires a fundamental change in the role of humans on the planet. This stewardship role requires deliberate and sustained action to become an integral, adaptive part of Earth System dynamics, creating feedbacks that keep the system on a Stabilized Earth pathway (*Alternative Stabilized Earth Pathway*).

We now explore this critical question in more detail by considering the relevant biogeophysical feedbacks (*Biogeophysical Feedbacks*) and the risk of tipping cascades (*Tipping Cascades*).

Biogeophysical Feedbacks. The trajectory of the Earth System is influenced by biogeophysical feedbacks within the system that can maintain it in a given state (negative feedbacks) and those that can amplify a perturbation and drive a transition to a different state (positive feedbacks). Some of the key negative feedbacks that could maintain the Earth System in Holocene-like conditions—notably, carbon uptake by land and ocean systems—are weakening relative to human forcing (19), increasing the risk that positive feedbacks could play an important role in determining the Earth System's trajectory. Table 1 summarizes carbon cycle feedbacks that could accelerate warming, while *SI Appendix*, Table S2 describes in detail a more complete set of biogeophysical feedbacks that can be triggered by forcing levels likely to be reached within the rest of the century.

Most of the feedbacks can show both continuous responses and tipping point behavior in which the feedback process becomes self-perpetuating after a critical threshold is crossed; subsystems exhibiting this behavior are often called "tipping elements" (17). The type of behavior—continuous response or tipping point/abrupt change—can depend on the magnitude or the rate of forcing, or both. Many feedbacks will show some gradual change before the tipping point is reached.

A few of the changes associated with the feedbacks are reversible on short timeframes of 50–100 years (e.g., change in Arctic sea ice extent with a warming or cooling of the climate; Antarctic sea ice may be less reversible because of heat accumulation in the Southern Ocean), but most changes are largely irreversible on timeframes that matter to contemporary societies (e.g., loss of permafrost carbon). A few of the feedbacks do not have apparent thresholds (e.g., change in the land and ocean physiological carbon sinks, such as increasing carbon uptake due

Table 1. **Carbon cycle feedbacks in the Earth System that could accelerate global warming**

Feedback	Strength of feedback by 2100,* °C	Refs. (*SI Appendix*, Table S2 has more details)
Permafrost thawing	0.09 (0.04–0.16)	20–23
Relative weakening of land and ocean physiological C sinks	0.25 (0.13–0.37)	24
Increased bacterial respiration in the ocean	0.02	25, 26
Amazon forest dieback	0.05 (0.03–0.11)	27
Boreal forest dieback	0.06 (0.02–0.10)	28
Total	0.47 (0.24–0.66)	

The strength of the feedback is estimated at 2100 for an ∼2 °C warming.
*The additional temperature rise (degrees Celsius) by 2100 arising from the feedback.

Reproduced from: W. Steffen et al., Trajectories of the Earth System in the Anthropocene, Proc. Natl. Acad. Sci. U. S. A., vol. 115, no. 33, pp. 8252–8259, 2018, doi: 10.1073/pnas.1810141115. Published under Creative Commons Attribution License 4.0 (CC BY NC ND).

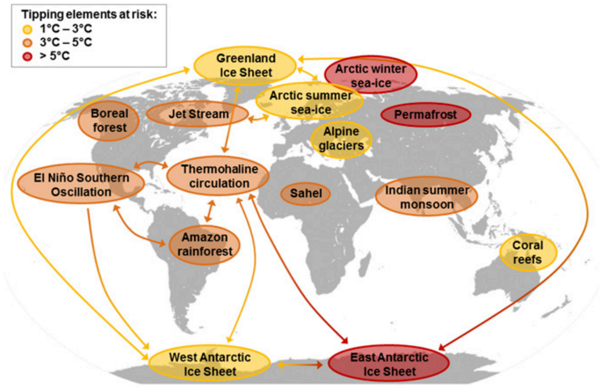


Fig. 3. Global map of potential tipping cascades. The individual tipping elements are color- coded according to estimated thresholds in global average surface temperature (tipping points) (12, 34). Arrows show the potential interactions among the tipping elements based on expert elicitation that could generate cascades. Note that, although the risk for tipping (loss of) the East Antarctic Ice Sheet is proposed at >5 °C, some marine-based sectors in East Antarctica may be vulnerable at lower temperatures (35–38).

to the CO_2 fertilization effect or decreasing uptake due to a decrease in rainfall). For some of the tipping elements, crossing the tipping point could trigger an abrupt, nonlinear response (e.g., conversion of large areas of the Amazon rainforest to a savanna or seasonally dry forest), while for others, crossing the tipping point would lead to a more gradual but self-perpetuating response (large-scale loss of permafrost). There could also be considerable lags after the crossing of a threshold, particularly for those tipping elements that involve the melting of large masses of ice. However, in some cases, ice loss can be very rapid when occurring as massive iceberg outbreaks (e.g., Heinrich Events).

For some feedback processes, the magnitude—and even the direction—depend on the rate of climate change. If the rate of climate change is small, the shift in biomes can track the change in temperature/moisture, and the biomes may shift gradually, potentially taking up carbon from the atmosphere as the climate warms and atmospheric CO_2 concentration increases. However, if the rate of climate change is too large or too fast, a tipping point can be crossed, and a rapid biome shift may occur via extensive disturbances (e.g., wildfires, insect attacks, droughts) that can abruptly remove an existing biome. In some terrestrial cases, such as widespread wildfires, there could be a pulse of carbon to the atmosphere, which if large enough, could influence the trajectory of the Earth System (29).

Varying response rates to a changing climate could lead to complex biosphere dynamics with implications for feedback processes. For example, delays in permafrost thawing would most likely delay the projected northward migration of boreal forests (30), while warming of the southern areas of these forests could result in their conversion to steppe grasslands of significantly lower carbon storage capacity. The overall result would be a positive feedback to the climate system.

The so-called "greening" of the planet, caused by enhanced plant growth due to increasing atmospheric CO_2 concentration (31), has increased the land carbon sink in recent decades (32). However, increasing atmospheric CO_2 raises temperature, and hotter leaves photosynthesize less well. Other feedbacks are also involved—for instance, warming the soil increases microbial respiration, releasing CO_2 back into the atmosphere.

Our analysis focuses on the strength of the feedback between now and 2100. However, several of the feedbacks that show negligible or very small magnitude by 2100 could nevertheless be triggered well before then, and they could eventually generate significant feedback strength over longer timeframes—centuries and even millennia—and thus, influence the long-term trajectory of the Earth System. These feedback processes include permafrost thawing, decomposition of ocean methane hydrates, increased marine bacterial respiration, and loss of polar ice sheets accompanied by a rise in sea levels and potential amplification of temperature rise through changes in ocean circulation (33).

Tipping Cascades. Fig. 3 shows a global map of some potential tipping cascades. The tipping elements fall into three clusters based on their estimated threshold temperature (12, 17, 39). Cascades could be formed when a rise in global temperature reaches the level of the lower-temperature cluster, activating tipping elements, such as loss of the Greenland Ice Sheet or Arctic sea ice. These tipping elements, along with some of the non-tipping element feedbacks (e.g., gradual weakening of land and ocean physiological carbon sinks), could push the global average temperature even higher, inducing tipping in mid- and higher-temperature clusters. For example, tipping (loss) of the Greenland Ice Sheet could trigger a critical transition in the Atlantic Meridional Ocean Circulation (AMOC), which could together, by causing sea-level rise and Southern Ocean heat accumulation, accelerate ice loss from the East Antarctic Ice Sheet (32, 40) on timescales of centuries (41).

Observations of past behavior support an important contribution of changes in ocean circulation to such feedback cascades. During previous glaciations, the climate system flickered between two states that seem to reflect changes in convective activity in the Nordic seas and changes in the activity of the AMOC. These variations caused typical temperature response patterns called the "bipolar seesaw" (42–44). During extremely cold conditions in the north, heat accumulated in the Southern Ocean, and Antarctica warmed. Eventually, the heat made its way north and generated subsurface warming that may have been instrumental in destabilizing the edges of the Northern Hemisphere ice sheets (45).

If Greenland and the West Antarctic Ice Sheet melt in the future, the freshening and cooling of nearby surface waters will have significant effects on the ocean circulation. While the probability of significant circulation changes is difficult to quantify, climate model simulations suggest that freshwater inputs compatible with current rates of Greenland melting are sufficient to have measurable effects on ocean temperature and circulation (46, 47). Sustained warming of the northern high latitudes as a result of this process could accelerate feedbacks or activate tipping elements in that region, such as permafrost degradation, loss of Arctic sea ice, and boreal forest dieback.

While this may seem to be an extreme scenario, it illustrates that a warming into the range of even the lower-temperature cluster (i.e., the Paris targets) could lead to tipping in the mid- and higher-temperature clusters via cascade effects. Based on this analysis of tipping cascades and taking a risk-averse approach, we suggest that a potential planetary threshold could occur at a temperature rise as low as ~2.0 °C above preindustrial (Fig. 1).

Alternative Stabilized Earth Pathway

If the world's societies want to avoid crossing a potential threshold that locks the Earth System into the Hothouse Earth pathway, then it is critical that they make deliberate decisions to avoid this risk

Reproduced from: W. Steffen et al., Trajectories of the Earth System in the Anthropocene, Proc. Natl. Acad. Sci. U. S. A., vol. 115, no. 33, pp. 8252–8259, 2018, doi: 10.1073/pnas.1810141115. Published under Creative Commons Attribution License 4.0 (CC BY NC ND).

and maintain the Earth System in Holocene-like conditions. This human-created pathway is represented in Figs. 1 and 2 by what we call Stabilized Earth (small loop at the bottom of Fig. 1, *Upper Right*), in which the Earth System is maintained in a state with a temperature rise no greater than 2 °C above preindustrial (a "super-Holocene" state) (11). Stabilized Earth would require deep cuts in greenhouse gas emissions, protection and enhancement of biosphere carbon sinks, efforts to remove CO_2 from the atmosphere, possibly solar radiation management, and adaptation to unavoidable impacts of the warming already occurring (48). The short broken red line beyond Stabilized Earth in Fig. 1, *Upper Right* represents a potential return to interglacial-like conditions in the longer term.

In essence, the Stabilized Earth pathway could be conceptualized as a regime of the Earth System in which humanity plays an active planetary stewardship role in maintaining a state intermediate between the glacial–interglacial limit cycle of the Late Quaternary and a Hothouse Earth (Fig. 2). We emphasize that Stabilized Earth is not an intrinsic state of the Earth System but rather, one in which humanity commits to a pathway of ongoing management of its relationship with the rest of the Earth System.

A critical issue is that, if a planetary threshold is crossed toward the Hothouse Earth pathway, accessing the Stabilized Earth pathway would become very difficult no matter what actions human societies might take. Beyond the threshold, positive (reinforcing) feedbacks within the Earth System—outside of human influence or control—could become the dominant driver of the system's pathway, as individual tipping elements create linked cascades through time and with rising temperature (Fig. 3). In other words, after the Earth System is committed to the Hothouse Earth pathway, the alternative Stabilized Earth pathway would very likely become inaccessible as illustrated in Fig. 2.

What Is at Stake? Hothouse Earth is likely to be uncontrollable and dangerous to many, particularly if we transition into it in only a century or two, and it poses severe risks for health, economies, political stability (12, 39, 49, 50) (especially for the most climate vulnerable), and ultimately, the habitability of the planet for humans.

Insights into the risks posed by the rapid climatic changes emerging in the Anthropocene can be obtained not only from contemporary observations (51–55) but also, from interactions in the past between human societies and regional and seasonal hydroclimate variability. This variability was often much more pronounced than global, longer-term Holocene variability (*SI Appendix*). Agricultural production and water supplies are especially vulnerable to changes in the hydroclimate, leading to hot/dry or cool/wet extremes. Societal declines, collapses, migrations/resettlements, reorganizations, and cultural changes were often associated with severe regional droughts and with the global megadrought at 4.2–3.9 thousand years before present, all occurring within the relative stability of the narrow global Holocene temperature range of approximately ±1 °C (56).

SI Appendix, Table S4 summarizes biomes and regional biosphere–physical climate subsystems critical for human wellbeing and the resultant risks if the Earth System follows a Hothouse Earth pathway. While most of these biomes or regional systems may be retained in a Stabilized Earth pathway, most or all of them would likely be substantially changed or degraded in a Hothouse Earth pathway, with serious challenges for the viability of human societies.

For example, agricultural systems are particularly vulnerable, because they are spatially organized around the relatively stable Holocene patterns of terrestrial primary productivity, which depend on a well-established and predictable spatial distribution of temperature and precipitation in relation to the location of fertile soils as well as on a particular atmospheric CO_2 concentration. Current understanding suggests that, while a Stabilized Earth pathway could result in an approximate balance between increases and decreases in regional production as human systems adapt, a Hothouse Earth trajectory will likely exceed the limits of adaptation and result in a substantial overall decrease in agricultural production, increased prices, and even more disparity between wealthy and poor countries (57).

The world's coastal zones, especially low-lying deltas and the adjacent coastal seas and ecosystems, are particularly important for human wellbeing. These areas are home to much of the world's population, most of the emerging megacities, and a significant amount of infrastructure vital for both national economies and international trade. A Hothouse Earth trajectory would almost certainly flood deltaic environments, increase the risk of damage from coastal storms, and eliminate coral reefs (and all of the benefits that they provide for societies) by the end of this century or earlier (58).

Human Feedbacks in the Earth System. In the dominant climate change narrative, humans are an external force driving change to the Earth System in a largely linear, deterministic way; the higher the forcing in terms of anthropogenic greenhouse gas emissions, the higher the global average temperature. However, our analysis argues that human societies and our activities need to be recast as an integral, interacting component of a complex, adaptive Earth System. This framing puts the focus not only on human system dynamics that reduce greenhouse gas emissions but also, on those that create or enhance negative feedbacks that reduce the risk that the Earth System will cross a planetary threshold and lock into a Hothouse Earth pathway.

Humanity's challenge then is to influence the dynamical properties of the Earth System in such a way that the emerging unstable conditions in the zone between the Holocene and a very hot state become a de facto stable intermediate state (Stabilized Earth) (Fig. 2). This requires that humans take deliberate, integral, and adaptive steps to reduce dangerous impacts on the Earth System, effectively monitoring and changing behavior to form feedback loops that stabilize this intermediate state.

There is much uncertainty and debate about how this can be done—technically, ethically, equitably, and economically—and there is no doubt that the normative, policy, and institutional aspects are highly challenging. However, societies could take a wide range of actions that constitute negative feedbacks, summarized in *SI Appendix*, Table S5, to steer the Earth System toward Stabilized Earth. Some of these actions are already altering emission trajectories. The negative feedback actions fall into three broad categories: (*i*) reducing greenhouse gas emissions, (*ii*) enhancing or creating carbon sinks (e.g., protecting and enhancing biosphere carbon sinks and creating new types of sinks) (59), and (*iii*) modifying Earth's energy balance (for example, via solar radiation management, although that particular feedback entails very large risks of destabilization or degradation of several key processes in the Earth System) (60, 61). While reducing emissions is a priority, much more could be done to reduce direct human pressures on critical biomes that contribute to the regulation of the state of the Earth System through carbon sinks and moisture feedbacks, such as the Amazon and boreal forests (Table 1), and to build much more effective stewardship of the marine and terrestrial biospheres in general.

The present dominant socioeconomic system, however, is based on high-carbon economic growth and exploitative resource use (9). Attempts to modify this system have met with some

Reproduced from: W. Steffen et al., Trajectories of the Earth System in the Anthropocene, Proc. Natl. Acad. Sci. U. S. A., vol. 115, no. 33, pp. 8252–8259, 2018, doi: 10.1073/pnas.1810141115. Published under Creative Commons Attribution License 4.0 (CC BY NC ND).

success locally but little success globally in reducing greenhouse gas emissions or building more effective stewardship of the biosphere. Incremental linear changes to the present socioeconomic system are not enough to stabilize the Earth System. Widespread, rapid, and fundamental transformations will likely be required to reduce the risk of crossing the threshold and locking in the Hothouse Earth pathway; these include changes in behavior, technology and innovation, governance, and values (48, 62, 63).

International efforts to reduce human impacts on the Earth System while improving wellbeing include the United Nations Sustainable Development Goals and the commitment in the Paris agreement to keep warming below 2 °C. These international governance initiatives are matched by carbon reduction commitments by countries, cities, businesses, and individuals (64–66) , but as yet, these are not enough to meet the Paris target. Enhanced ambition will need new collectively shared values, principles, and frameworks as well as education to support such changes (67, 68). In essence, effective Earth System stewardship is an essential precondition for the prosperous development of human societies in a Stabilized Earth pathway (69, 70).

In addition to institutional and social innovation at the global governance level, changes in demographics, consumption, behavior, attitudes, education, institutions, and socially embedded technologies are all important to maximize the chances of achieving a Stabilized Earth pathway (71). Many of the needed shifts may take decades to have a globally aggregated impact (*SI Appendix*, Table S5), but there are indications that society may be reaching some important societal tipping points. For example, there has been relatively rapid progress toward slowing or reversing population growth through declining fertility resulting from the empowerment of women, access to birth control technologies, expansion of educational opportunities, and rising income levels (72, 73). These demographic changes must be complemented by sustainable per capita consumption patterns, especially among the higher per capita consumers. Some changes in consumer behavior have been observed (74, 75), and opportunities for consequent major transitions in social norms over broad scales may arise (76). Technological innovation is contributing to more rapid decarbonization and the possibility for removing CO_2 from the atmosphere (48).

Ultimately, the transformations necessary to achieve the Stabilized Earth pathway require a fundamental reorientation and restructuring of national and international institutions toward more effective governance at the Earth System level (77), with a much stronger emphasis on planetary concerns in economic governance, global trade, investments and finance, and technological development (78).

Building Resilience in a Rapidly Changing Earth System. Even if a Stabilized Earth pathway is achieved, humanity will face a turbulent road of rapid and profound changes and uncertainties on route to it—politically, socially, and environmentally—that challenge the resilience of human societies (79–82). Stabilized Earth will likely be warmer than any other time over the last 800,000 years at least (83) (that is, warmer than at any other time in which fully modern humans have existed).

In addition, the Stabilized Earth trajectory will almost surely be characterized by the activation of some tipping elements (*Tipping Cascades* and Fig. 3) and by nonlinear dynamics and abrupt shifts at the level of critical biomes that support humanity (*SI Appendix*, Table S4). Current rates of change of important features of the Earth System already match or exceed those of abrupt geophysical events in the past (*SI Appendix*). With these trends likely to continue for the next several decades at least, the contemporary way of guiding development founded on theories, tools, and beliefs of gradual or incremental change, with a focus on economy efficiency, will likely not be adequate to cope with this trajectory. Thus, in addition to adaptation, increasing resilience will become a key strategy for navigating the future.

Generic resilience-building strategies include developing insurance, buffers, redundancy, diversity, and other features of resilience that are critical for transforming human systems in the face of warming and possible surprise associated with tipping points (84). Features of such a strategy include (*i*) maintenance of diversity, modularity, and redundancy; (*ii*) management of connectivity, openness, slow variables, and feedbacks; (*iii*) understanding social–ecological systems as complex adaptive systems, especially at the level of the Earth System as a whole (85); (*iv*) encouraging learning and experimentation; and (*v*) broadening of participation and building of trust to promote polycentric governance systems (86, 87).

Conclusions

Our systems approach, focusing on feedbacks, tipping points, and nonlinear dynamics, has addressed the four questions posed in the Introduction.

Our analysis suggests that the Earth System may be approaching a planetary threshold that could lock in a continuing rapid pathway toward much hotter conditions—Hothouse Earth. This pathway would be propelled by strong, intrinsic, biogeophysical feedbacks difficult to influence by human actions, a pathway that could not be reversed, steered, or substantially slowed.

Where such a threshold might be is uncertain, but it could be only decades ahead at a temperature rise of ~2.0 °C above preindustrial, and thus, it could be within the range of the Paris Accord temperature targets.

The impacts of a Hothouse Earth pathway on human societies would likely be massive, sometimes abrupt, and undoubtedly disruptive.

Avoiding this threshold by creating a Stabilized Earth pathway can only be achieved and maintained by a coordinated, deliberate effort by human societies to manage our relationship with the rest of the Earth System, recognizing that humanity is an integral, interacting component of the system. Humanity is now facing the need for critical decisions and actions that could influence our future for centuries, if not millennia (88).

How credible is this analysis? There is significant evidence from a number of sources that the risk of a planetary threshold and thus, the need to create a divergent pathway should be taken seriously:

First, the complex system behavior of the Earth System in the Late Quaternary is well-documented and understood. The two bounding states of the system—glacial and interglacial—are reasonably well-defined, the ca. 100,000-years periodicity of the limit cycle is established, and internal (carbon cycle and ice albedo feedbacks) and external (changes in insolation caused by changes in Earth's orbital parameters) driving processes are generally well-known. Furthermore, we know with high confidence that the progressive disintegration of ice sheets and the transgression of other tipping elements are difficult to reverse after critical levels of warming are reached.

Second, insights from Earth's recent geological past (*SI Appendix*) suggest that conditions consistent with the Hothouse Earth pathway are accessible with levels of atmospheric CO_2 concentration and temperature rise either already realized or projected for this century (*SI Appendix*, Table S1).

Reproduced from: W. Steffen et al., Trajectories of the Earth System in the Anthropocene, Proc. Natl. Acad. Sci. U. S. A., vol. 115, no. 33, pp. 8252–8259, 2018, doi: 10.1073/pnas.1810141115. Published under Creative Commons Attribution License 4.0 (CC BY NC ND).

Third, the tipping elements and feedback processes that operated over Quaternary glacial–interglacial cycles are the same as several of those proposed as critical for the future trajectory of the Earth System (*Biogeophysical Feedbacks*, *Tipping Cascades*, Fig. 3, Table 1, and *SI Appendix*, Table S2).

Fourth, contemporary observations (29, 38) (*SI Appendix*) of tipping element behavior at an observed temperature anomaly of about 1 °C above preindustrial suggest that some of these elements are vulnerable to tipping within just a 1 °C to 3 °C increase in global temperature, with many more of them vulnerable at higher temperatures (*Biogeophysical Feedbacks* and *Tipping Cascades*) (12, 17, 39). This suggests that the risk of tipping cascades could be significant at a 2 °C temperature rise and could increase sharply beyond that point. We argue that a planetary threshold in the Earth System could exist at a temperature rise as low as 2 °C above preindustrial.

The Stabilized Earth trajectory requires deliberate management of humanity's relationship with the rest of the Earth System if the world is to avoid crossing a planetary threshold. We suggest that a deep transformation based on a fundamental reorientation of human values, equity, behavior, institutions, economies, and technologies is required. Even so, the pathway toward Stabilized Earth will involve considerable changes to the structure and functioning of the Earth System, suggesting that resilience-building strategies be given much higher priority than at present in decision making. Some signs are emerging that societies are initiating some of the necessary transformations. However, these transformations are still in initial stages, and the social/political tipping points that definitively move the current trajectory away from Hothouse Earth have not yet been crossed, while the door to the Stabilized Earth pathway may be rapidly closing.

Our initial analysis here needs to be underpinned by more in-depth, quantitative Earth System analysis and modeling studies to address three critical questions. (*i*) Is humanity at risk for pushing the system across a planetary threshold and irreversibly down a Hothouse Earth pathway? (*ii*) What other pathways might be possible in the complex stability landscape of the Earth System, and what risks might they entail? (*iii*) What planetary stewardship strategies are required to maintain the Earth System in a manageable Stabilized Earth state?

Acknowledgments

We thank the three reviewers for their comments on the first version of the manuscript and two of the reviewers for further comments on a revised version of the manuscript. These comments were very helpful in the revisions. We thank a member of the PNAS editorial board for a comprehensive and very helpful review. W.S. and C.P.S. are members of the Anthropocene Working Group. W.S., J.R., K.R., S.E.C., J.F.D., I.F., S.J.L., R.W. and H.J.S. are members of the Planetary Boundaries Research Network PB.net and the Earth League's EarthDoc Programme supported by the Stordalen Foundation. T.M.L. was supported by a Royal Society Wolfson Research Merit Award and the European Union Framework Programme 7 Project HELIX. C.F. was supported by the Erling–Persson Family Foundation. The participation of D.L. was supported by the Haury Program in Environment and Social Justice and National Science Foundation (USA) Decadal and Regional Climate Prediction using Earth System Models Grant 1243125. S.E.C. was supported in part by Swedish Research Council Formas Grant 2012-742. J.F.D. and R.W. were supported by Leibniz Association Project DOMINOES. S.J.L. receives funding from Formas Grant 2014-589. This paper is a contribution to European Research Council Advanced Grant 2016, Earth Resilience in the Anthropocene Project 743080.

1 Crutzen PJ (2002) Geology of mankind. *Nature* 415:23.
2 Steffen W, Broadgate W, Deutsch L, Gaffney O, Ludwig C (2015) The trajectory of the Anthropocene: The great acceleration. *Anthropocene Rev* 2:81–98.
3 Waters CN, et al. (2016) The Anthropocene is functionally and stratigraphically distinct from the Holocene. *Science* 351:aad2622.
4 Malm A, Hornborg A (2014) The geology of mankind? A critique of the Anthropocene narrative. *Anthropocene Rev* 1:62–69.
5 Donges JF, et al. (2017) Closing the loop: Reconnecting human dynamics to Earth System science. *Anthropocene Rev* 4:151–157.
6 Levin SA (2003) Complex adaptive systems: Exploring the known, the unknown and the unknowable. *Bull Am Math Soc* 40:3–20.
7 Past Interglacial Working Group of PAGES (2016) Interglacials of the last 800,000 years. *Rev Geophys* 54:162–219.
8 Williams M, et al. (2015) The Anthropocene biosphere. *Anthropocene Rev* 2:196–219.
9 McNeill JR, Engelke P (2016) *The Great Acceleration* (Harvard Univ Press, Cambridge, MA).
10 Hawkins E, et al. (2017) Estimating changes in global temperature since the pre-industrial period. *Bull Am Meteorol Soc* 98:1841–1856.
11 Ganopolski A, Winkelmann R, Schellnhuber HJ (2016) Critical insolation-CO_2 relation for diagnosing past and future glacial inception. *Nature* 529:200–203.
12 Schellnhuber HJ, Rahmstorf S, Winkelmann R (2016) Why the right climate target was agreed in Paris. *Nat Clim Change* 6:649–653.
13 Schellnhuber HJ (1999) 'Earth system' analysis and the second Copernican revolution. *Nature* 402(Suppl):C19–C23.
14 IPCC (2013) Summary for policymakers. *Climate Change 2013: The Physical Science Basis, Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Stocker TF, et al. (Cambridge Univ Press, Cambridge, UK), pp 3–29.
15 Drijfhout S, et al. (2015) Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models. *Proc Natl Acad Sci USA* 112:E5777–E5786.
16 Stocker TF, et al. (2013) Technical summary. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Stocker TF, et al. (Cambridge Univ Press, Cambridge, UK).
17 Lenton TM, et al. (2008) Tipping elements in the Earth's climate system. *Proc Natl Acad Sci USA* 105:1786–1793.
18 Scheffer M (2009) *Critical Transitions in Nature and Society* (Princeton Univ Press, Princeton).
19 Raupach MR, et al. (2014) The declining uptake rate of atmospheric CO_2 by land and ocean sinks. *Biogeosciences* 11:3453–3475.
20 Schaefer K, Lantuit H, Romanovsky VE, Schuur EAG, Witt R (2014) The impact of the permafrost carbon feedback on global climate. *Environ Res Lett* 9:085003.
21 Schneider von Deimling T, et al. (2015) Observation-based modelling of permafrost carbon fluxes with accounting for deep carbon deposits and thermokarst activity. *Biogeosciences* 12:3469–3488.
22 Koven CD, et al. (2015) A simplified, data-constrained approach to estimate the permafrost carbon-climate feedback. *Philos Trans A Math Phys Eng Sci* 373:20140423.
23 Chadburn SE, et al. (2017) An observation-based constraint on permafrost loss as a function of global warming. *Nat Clim Change* 7:340–344.
24 Ciais P, et al. (2013) Carbon and other biogeochemical cycles. *Climate Change 2013: The Physical Science Basis, Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Stocker TF, et al. (Cambridge Univ Press, Cambridge, UK), pp 465–570.
25 Segschneider J, Bendtsen J (2013) Temperature-dependent remineralization in a warming ocean increases surface pCO_2 through changes in marine ecosystem composition. *Global Biogeochem Cycles* 27:1214–1225.
26 Bendtsen J, Hilligsøe KM, Hansen J, Richardson K (2015) Analysis of remineralisation, lability, temperature sensitivity and structural composition of organic matter from the upper ocean. *Prog Oceanogr* 130:125–145.
27 Jones C, Lowe J, Liddicoat S, Betts R (2009) Committed terrestrial ecosystem changes due to climate change. *Nat Geosci* 2:484–487.
28 Kurz WA, Apps MJ (1999) A 70-year retrospective analysis of carbon fluxes in the Canadian forest sector. *Ecol Appl* 9:526–547.
29 Lewis SL, Brando PM, Phillips OL, van der Heijden GMF, Nepstad D (2011) The 2010 Amazon drought. *Science* 331:554.
30 Herzschuh U, et al. (2016) Glacial legacies on interglacial vegetation at the Pliocene-Pleistocene transition in NE Asia. *Nature Commun* 7:11967.

Reproduced from: W. Steffen et al., Trajectories of the Earth System in the Anthropocene, Proc. Natl. Acad. Sci. U. S. A., vol. 115, no. 33, pp. 8252–8259, 2018, doi: 10.1073/pnas.1810141115. Published under Creative Commons Attribution License 4.0 (CC BY NC ND).

31 Mao J, et al. (2016) Human-induced greening of the northern extratropical land surface. *Nat Clim Change* 6:959–963.

32 Keenan TF, et al. (2016) Recent pause in the growth rate of atmospheric CO_2 due to enhanced terrestrial carbon uptake. *Nature Commun* 7:13428, and erratum (2017) 8:16137.

33 Hansen J, et al. (2016) Ice melt, sea level rise and superstorms: Evidence from paleoclimatedata, climate modeling, and modern observations that 2 °C global warming could be dangerous. *Atmos Chem Phys* 16:3761–3812.

34 Kriegler E, Hall JW, Held H, Dawson R, Schellnhuber HJ (2009) Imprecise probability assessment of tipping points in the climate system. *Proc Natl Acad Sci USA* 106:5041–5046.

35 Pollard D, DeConto RM (2009) Modelling West Antarctic ice sheet growth and collapse through the past five million years. *Nature* 458:329–332.

36 Pollard D, DeConto RM, Alley RB (2015) Potential Antarctic Ice Sheet retreat driven by hydrofracturing and ice cliff failure. *Earth Planet Sci Lett* 412:112–121.

37 DeConto RM, Pollard D (2016) Contribution of Antarctica to past and future sea-level rise. *Nature* 531:591–597.

38 Rintoul SR, et al. (2016) Ocean heat drives rapid basal melt of the Totten Ice Shelf. *Sci Adv* 2:e1601610.

39 US Department of Defense (2015) National security implications of climate-related risks and a changing climate. Available at archive.defense.gov/pubs/150724-congressional-report-on-national-implications-of-climate-change.pdf?source=govdelivery. Accessed February 7, 2018.

40 Mengel M, Levermann A (2014) Ice plug prevents irreversible discharge from East Antarctica. *Nat Clim Change* 4:451–455.

41 Armour KC, et al. (2016) Southern Ocean warming delayed by circumpolar upwelling and equatorward transport. *Nat Geosci* 9:549–554.

42 Stocker TF, Johnsen SJ (2003) A minimum thermodynamic model for the bipolar seesaw. *Paleoceanography* 18:1087.

43 Rahmstorf S (2002) Ocean circulation and climate during the past 120,000 years. *Nature* 419:207–214.

44 Hemming SR (2004) Heinrich events: Massive late Pleistocene detritus layers of the North Atlantic and their global climate imprint. *Rev Geophys* 42:1–43.

45 Alvarez-Solas J, et al. (2010) Link between ocean temperature and iceberg discharge during Heinrich events. *Nat Geosci* 3:122–126.

46 Stouffer RJ, et al. (2006) Investigating the causes of the response of the thermohaline circulation to past and future climate changes. *J Clim* 19:1365–1387.

47 Swingedow D, et al. (2013) Decadal fingerprints of freshwater discharge around Greenland in a multi-model ensemble. *Clim Dyn* 41:695–720.

48 Rockström J, et al. (2017) A roadmap for rapid decarbonization. *Science* 355:1269–1271.

49 Schleussner C-F, Donges JF, Donner RV, Schellnhuber HJ (2016) Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries. *Proc Natl Acad Sci USA* 113:9216–9221.

50 McMichael AJ, et al., eds (2003) *Climate Change and Human Health: Risks and Responses* (WHO, Geneva).

51 Udmale PD, et al. (2015) How did the 2012 drought affect rural livelihoods in vulnerable areas? Empirical evidence from India. *Int J Disaster Risk Reduct* 13:454–469.

52 Maldonado JK, Shearer C, Bronen R, Peterson K, Lazrus H (2013) The impact of climate change on tribal communities in the US: Displacement, relocation, and human rights. *Clim Change* 120:601–614.

53 Warner K, Afifi T (2014) Where the rain falls: Evidence from 8 countries on how vulnerable households use migration to manage the risk of rainfall variability and food insecurity. *Clim Dev* 6:1–17.

54 Cheung WW, Watson R, Pauly D (2013) Signature of ocean warming in global fisheries catch. *Nature* 497:365–368.

55 Nakano K (2017) Screening of climatic impacts on a country's international supply chains: Japan as a case study. *Mitig Adapt Strategies Glob Change* 22:651–667.

56 Latorre C, Wilmshurst J, von Gunten L, eds (2016) Climate change and cultural evolution. *PAGES (Past Global Changes) Magazine* 24:1–32.

57 IPCC (2014) Summary for policymakers. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Field CB, et al. (Cambridge Univ Press, Cambridge, UK), pp 1–32.

58 Schleussner C-F, et al. (2016) Science and policy characteristics of the Paris Agreement temperature goal. *Nat Clim Change* 6:827–835.

59 Griscom BW, et al. (2017) Natural climate solutions. *Proc Natl Acad Sci USA* 114:11645–11650.

60 Barrett S, et al. (2014) Climate engineering reconsidered. *Nat Clim Change* 4:527–529.

61 Mathesius S, Hofmann M, Calderia K, Schellnhuber HJ (2015) Long-term response of oceans to CO_2 removal from the atmosphere. *Nat Clim Change* 5:1107–1113.

62 Geels FW, Sovacool BK, Schwanen T, Sorrell S (2017) Sociotechnical transitions for deep decarbonization. *Science* 357:1242–1244.

63 O'Brien K (2018) Is the 1.5 °C target possible? Exploring the three spheres of transformation. *Curr Opin Environ Sustain* 31:153–160.

64 Young OR, et al. (2006) The globalization of socioecological systems: An agenda for scientific research. *Glob Environ Change* 16:304–316.

65 Adger NW, Eakin H, Winkels A (2009) Nested and teleconnected vulnerabilities to environmental change. *Front Ecol Environ* 7:150–157.

66 UN General Assembly (2015) *Transforming Our World: The 2030 Agenda for Sustainable Development*, A/RES/70/1. Available at https://sustainabledevelopment.un.org/content/documents/21252030%20Agenda%20for%20Sustainable%20Development%20web.pdf. Accessed July 18, 2018.

67 Wals AE, Brody M, Dillon J, Stevenson RB (2014) Science education. Convergence between science and environmental education. *Science* 344:583–584.

68 O'Brien K, et al. (2013) You say you want a revolution? Transforming education and capacity building in response to global change. *Environ Sci Policy* 28:48–59.

69 Chapin FS, III, et al. (2011) Earth stewardship: A strategy for social–ecological transformation to reverse planetary degradation. *J Environ Stud Sci* 1:44–53.

70 Folke C, Biggs R, Norström AV, Reyers B, Rockström J (2016) Social-ecological resilience and biosphere-based sustainability science. *Ecol Soc* 21:41.

71 Westley F, et al. (2011) Tipping toward sustainability: Emerging pathways of transformation. *Ambio* 40:762–780.

72 Lutz W, Muttarak R, Striessnig E (2014) Environment and development. Universal education is key to enhanced climate adaptation. *Science* 346:1061–1062.

73 Bongaarts J (2016) Development: Slow down population growth. *Nature* 530:409–412.

74 Defila R, Di Giulio A, Kaufmann-Hayoz R, eds (2012) *The Nature of Sustainable Consumption and How to Achieve It: Results from the Focal Topic "From Knowledge to Action–New Paths Towards Sustainable Consumption"* (Oakum, Munich).

75 Cohen MJ, Szejnwald Brown H, Vergragt P, eds (2013) *Innovations in Sustainable Consumption: New Economics, Socio-Technical Transitions and Social Practices* (Edward Elgar, Cheltenham, UK).

76 Nyborg K, et al. (2016) Social norms as solutions. *Science* 354:42–43.

77 Biermann F, et al. (2012) Science and government. Navigating the anthropocene: Improving Earth system governance. *Science* 335:1306–1307.

78 Galaz V (2014) *Global Environmental Governance, Technology and Politics: The Anthropocene Gap* (Edward Elgar, Cheltenham, UK).

79 Peters DPC, et al. (2004) Cross-scale interactions, nonlinearities, and forecasting catastrophic events. *Proc Natl Acad Sci USA* 101:15130–15135.

80 Walker B, et al. (2009) Environment. Looming global-scale failures and missing institutions. *Science* 325:1345–1346.

81 Hansen J, Sato M, Ruedy R (2012) Perception of climate change. *Proc Natl Acad Sci USA* 109:E2415–E2423.

82 Galaz V, et al. (2017) Global governance dimensions of globally networked risks: The state of the art in social science research. *Risks Hazards Crisis Public Policy* 8:4–27.

83 Augustin L, et al.; EPICA community members (2004) Eight glacial cycles from an Antarctic ice core. *Nature* 429:623–628.

84 Polasky S, Carpenter SR, Folke C, Keeler B (2011) Decision-making under great uncertainty: Environmental management in an era of global change. *Trends Ecol Evol* 26:398–404.

85 Capra F, Luisi PL (2014) *The Systems View of Life; A Unifying Vision* (Cambridge Univ Press, Cambridge, UK).

86 Carpenter SR, et al. (2012) General resilience to cope with extreme events. *Sustainability* 4:3248–3259.

87 Biggs R, et al. (2012) Toward principles for enhancing the resilience of ecosystem services. *Annu Rev Environ Resour* 37:421–448.

88 Figueres C, et al. (2017) Three years to safeguard our climate. *Nature* 546:593–595.

Earth Syst. Dynam., 8, 255–264, 2017
www.earth-syst-dynam.net/8/255/2017/
doi:10.5194/esd-8-255-2017

Earth System Dynamics
Open Access

Sustainable use of renewable resources in a stylized social–ecological network model under heterogeneous resource distribution

Wolfram Barfuss[1,3], **Jonathan F. Donges**[1,4], **Marc Wiedermann**[2,3], **and Wolfgang Lucht**[1,5,6]

[1]Earth System Analysis, Potsdam Institute for Climate Impact Research, Telegrafenberg A31, 14473 Potsdam, Germany
[2]Transdisciplinary Concepts & Methods, Potsdam Institute for Climate Impact Research, Telegrafenberg A31, 14473 Potsdam, Germany
[3]Department of Physics, Humboldt University, Newtonstraße 15, 12489 Berlin, Germany
[4]Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden
[5]Department of Geography, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany
[6]Integrative Research Institute on Transformations of Human-Environment Systems, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany

Correspondence to: Wolfram Barfuss (barfuss@pik-potsdam.de)

Received: 13 April 2016 – Discussion started: 30 May 2016
Revised: 3 February 2017 – Accepted: 24 February 2017 – Published: 11 April 2017

Abstract. Human societies depend on the resources ecosystems provide. Particularly since the last century, human activities have transformed the relationship between nature and society at a global scale. We study this coevolutionary relationship by utilizing a stylized model of private resource use and social learning on an adaptive network. The latter process is based on two social key dynamics beyond economic paradigms: boundedly rational imitation of resource use strategies and homophily in the formation of social network ties. The private and logistically growing resources are harvested with either a sustainable (small) or non-sustainable (large) effort. We show that these social processes can have a profound influence on the environmental state, such as determining whether the private renewable resources collapse from overuse or not. Additionally, we demonstrate that heterogeneously distributed regional resource capacities shift the critical social parameters where this resource extraction system collapses. We make these points to argue that, in more advanced coevolutionary models of the planetary social–ecological system, such socio-cultural phenomena as well as regional resource heterogeneities should receive attention in addition to the processes represented in established Earth system and integrated assessment models.

1 Resource use in social–ecological systems

Whether, when and how human usage of biophysical resources meets limits that produce feedbacks onto social functioning has a long history of controversial discussion (Malthus, 1798; Meadows et al., 1972; Rockström et al., 2009). Especially in the last century, human activities have changed the relationship between nature and society at the global scale (Crutzen, 2002; Steffen et al., 2007, 2015a), making them mutually interdependent in an unprecedented manner and the question of their joint dynamics urgent. Social and ecological systems should therefore be studied not only in isolation but also as interlinked social–ecological systems (Berkes and Folke, 1998). Here, we contribute to this debate by investigating properties of a stylized social system that cause the linked resource use system to either collapse or remain viable. Such a perspective also has important implications for the mathematical modeling of interdependent, global human–environment interactions (Verburg et al., 2016; van Vuuren et al., 2016). Typically, in present-

Published by Copernicus Publications on behalf of the European Geosciences Union.

day analysis the Earth system is either modeled from a purely biophysical point of view (Claussen et al., 2002) or from a biophysical–economic one (van Vuuren et al., 2012), depending on the scope of the research question. However, both approaches do not take into account social dynamics beyond macroeconomic paradigms.

Here, we conceptually explore avenues for a third strand of global modeling, next to the biophysical and biophysical–economic one, also incorporating socio-cultural dynamics. Founded on a genuinely social–ecological perspective, we term these "World–Earth" system models to emphasize the free coevolution of the social and ecological components (Schellnhuber, 1998, 1999). While sophisticated models of this type are not yet available, the literature contains various modeling studies that incorporate potentially important features such as static interaction networks (Chung et al., 2013; Sugiarto et al., 2015) to depict stylized social dynamics (Holme and Newman, 2006; Auer et al., 2015), tele-coupling effects in a globalized society interacting through social networks (Janssen et al., 2006; Bodin and Tengö, 2012), social–ecological regime shifts (Scheffer et al., 2001; Lade et al., 2013) and (social) tipping elements (Schellnhuber, 2009; Bentley et al., 2014), structural reorganization occurring on adaptive social networks (Gross and Blasius, 2008; Snijders et al., 2010; Sayama et al., 2013; Schleussner et al., 2016) or structural transformations (Lade et al., 2017) and cultural preference dynamics due to traits such as imitation (Traulsen et al., 2010) or homophily (McPherson et al., 2001; Centola et al., 2007).

We set out a simple model (see Sect. 2) to demonstrate that social network interactions, imitation and homophily may have a profound influence on the environmental state, such as determining whether a collection of private renewable resources collapses from overuse or not. We argue that more elaborate and sophisticated implementations of such social phenomena should receive attention in the future development of global system models, supplementing already established Earth system and integrated assessment models, neither of which at present include them.

As a particular case study for our model we examine the effect of heterogeneously distributed resources. This is important since in the real-world agents do have access to different amounts of biophysical resources. Our study examines under which combinations of parameters characterizing a social learning network process does the model converge to a sustainable regime for different degrees of resource access heterogeneity. Parameters governing social learning dynamics are, on the one hand, a homophily parameter ϕ, addressing the propensity of nodes to establish interactions with nodes of the same kind (see Sect. 2 for a detailed model description). On the other hand, the timescale of social interaction τ quantifies the average time for social updates on the network. We purposely do not model any form of individual learning of the agents with regard to the best harvesting strategy to emphasize the effects of the described social learning process. For homogeneous resource access (Wiedermann et al., 2015), one already observes a threshold in the parameter space of the model from non-sustainable to sustainable regimes at certain critical values ϕ_c and τ_c. Since the concrete heterogeneous resource distribution is often unknown, we show systematically how an increasing heterogeneity – starting from an almost homogeneous distribution – affects the critical transition parameters ϕ_c and τ_c. Additionally we show that in our stylized model a heavy-tailed resource distribution in comparison to a non-heavy-tailed distribution changes the model's behavior considerably. This is important as real-world resource data suggest that access to biophysical resources may indeed be distributed with heavy tails.

2 Model description

The intention behind our model design is not to closely follow any specific real-world setting but to explore the coevolution of socio-cultural dynamics with ecological dynamics. On a conceptual level, human–environment interactions are happening either in a common-pool or private-pool setting. Common-pool dilemmas have been studied extensively in the past (Hardin, 1968; Tavoni et al., 2012; Ostrom, 2015). Here, agents can retrieve information on another agent's harvesting strategy either via the ecological subsystem, i.e., the common pool, itself or via purely social interactions. In order to specifically focus on the latter of the two as an important domain of processes, we eliminate any transfer of information via the ecological system and discard a common-pool setting in favor of individual and private resource stocks per agent. Wiedermann et al. (2015) introduced a model for such a setting for the special case of homogeneously distributed private resources, revealed transitions and distinct regimes in its parameter space, and provided analytical approximations of its dynamics. Here, we adjust this setting for the more general case of an inhomogeneous resource distribution. An overview of the model is provided in Fig. 1.

2.1 A stylized anthroposphere

The social learning (Bandura, 1977) process takes place in a network initialized as a random graph G (Erdös and Rényi, 1960) with nodes labeled by integer number $i = 1, \ldots, N$ that represent social agents. It is based on two theoretical paradigms: (i) agents either change their strategy through boundedly rational imitation (Traulsen et al., 2010; Bahar et al., 2014) or (ii) adapt their local network structure by rewiring to other nodes with similar behavior (homophily, McPherson et al., 2001; Centola et al., 2007). In order to integrate this discrete update process (Holme and Newman, 2006; Zanette and Gil, 2006) with the continuous evolution of the resource stocks, social update times t_i are assigned to the agents as generated by a Poisson process with an expo-

Reproduced from: W. Barfuss et al., Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dyn., vol. 8, no. 2, pp. 255–264, 2017, doi: 10.5194/esd-8-255-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

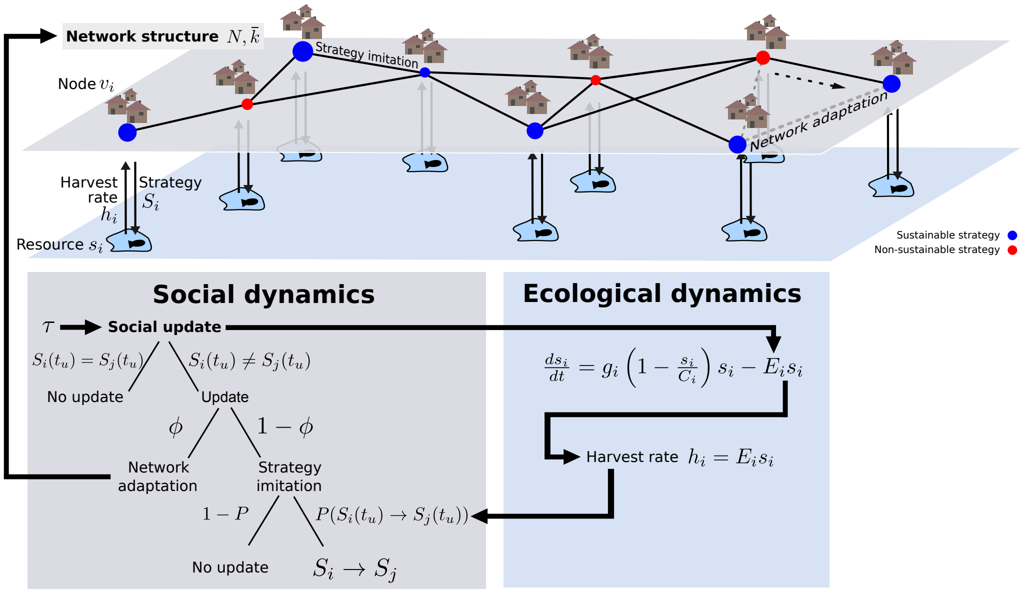


Figure 1. Illustration of our stylized social–ecological model. As the ecological subprocess the agents harvest their private logistically growing renewable resource with either a sustainable (blue) or non-sustainable (red) strategy. The social subprocess follows the logics of strategy imitation due to comparisons of harvest rates and of social network adaptation due to homophily. The social update times are generated by a Poisson process with average inter-event time τ.

nential distribution,

$$p(\Delta t_i; \tau) = \frac{1}{\tau} \exp\left(\frac{-\Delta t_i}{\tau}\right), \tag{1}$$

of waiting times Δt_i, where the parameter τ gives the expected waiting time.

Thus, agent v_i with the lowest update time in the queue performs the social update process accordingly:

- (1) If the degree of agent v_i is zero (i.e., v_i has no neighbors), move to (3); otherwise choose a neighbor v_j of v_i at random.

- (2) If v_j and v_i employ the same harvesting strategy $S_i = S_j$ (either sustainable or non-sustainable; see below), move to (3). Otherwise, move to (2.1).

- (2.1) With rewiring probability ϕ disconnect v_j from v_i and connect v_i to a randomly chosen agent v_k that employs the same strategy.

- (2.2) If (2.1) was not chosen, change the strategy of v_i to the one of v_j according to the sigmoidal imitation probability function

  $$P(S_i \rightarrow S_j) = \frac{1}{2}\left(\tanh\left(\gamma\left[h_j(t) - h_i(t)\right]\right) + 1\right). \tag{2}$$

  Hence, the greater the harvest rate h_j (see below) of v_j with respect to the harvest rate h_i of v_i, the more likely agent v_i is to change its strategy to the one of agent v_j. Agents only consider their current yields when formulating their next harvesting strategy. This assumption reflects boundedly rational behavior in the form of the agent's limited knowledge of their own and their neighbors' ecosystems. The parameter γ controls the slope of the imitation probability function (Eq. 2) – i.e., for $\gamma \rightarrow \infty$ node v_i would always imitate agent v_j's strategy if $h_j(t) > h_i(t)$, while for $\gamma \rightarrow 0$ the imitation probability tends to $1/2$ and is independent of the agents' harvest rates. Therefore, one can interpret γ as an imitation tendency parameter. In fact, Traulsen et al. (2010) found this sigmoidal shape of imitation probability in a behavioral experiment.

- (3) For the next update, another waiting time is drawn from the exponential distribution (Eq. 1) and added to the update time of node v_i.

2.2 A stylized ecosphere

2.2.1 Private resource dynamics

The ecological module of our model consists of private renewable resources each following a logistic growth function, which is chosen as one of the simplest and most commonly used models of renewable resource dynamics in a constrained environment (Brander and Taylor, 1998; Keeling, 2000; Perman et al., 2003). Additionally, a harvest rate

Reproduced from: W. Barfuss et al., Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dyn., vol. 8, no. 2, pp. 255–264, 2017, doi: 10.5194/esd-8-255-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

$h_i = E_i s_i$ is subtracted from the rate of change of the resource stock s_i. E_i denotes the effort of agent v_i. Thus, the dynamics of the ith resource are given by

$$\frac{\mathrm{d}s_i}{\mathrm{d}t} = g_i \left(1 - \frac{s_i}{C_i}\right) s_i - E_i s_i. \qquad (3)$$

Here, g_i denotes the growth rate and C_i the carrying capacity of the ith resource stock. The strategy S_i of agent v_i can either be sustainable ($S_i = 1$), resulting in an effort $E_{i,s} = \frac{g_i}{2}$, or non-sustainable ($S_i = 0$) with an effort $E_{i,n} = \frac{3g_i}{2}$. These efforts have been chosen such that the sustainable strategy coincides with the maximum sustainable yield, whereas the non-sustainable strategy leads to the full depletion of the resource stock and, consequently, no harvest at all in the long term. Note that $E_{i,\mathrm{n}}$ and $E_{i,\mathrm{s}}$ are symmetrically separated from the critical effort $E_{i,\mathrm{c}} = g_i$. The latter is defined such that, for positive efforts below $E_{i,\mathrm{c}}$, the resource stock converges to a non-zero stationary state, whereas for efforts above $E_{i,\mathrm{c}}$ the resource stock collapses and converges to zero. When in interplay with the social update process, Eq. (3) is used as its analytically derived definite integral, which circumvents the need for any numerical integration methods.

2.2.2 Resource heterogeneity

Heterogeneous access to resources is operationalized by randomly distributing the resource capacities C_i according to a prescribed probability density function. For this purpose, we examine the lognormal distribution

$$\ln\mathcal{N}(C;\mu,\sigma) = \frac{1}{C\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln C - \mu)^2}{2\sigma^2}\right], \quad C > 0, \qquad (4)$$

with parameters μ and σ (not to be confused with the standard deviation of C). It is derived from the normal distribution: a positive random variable is lognormally distributed if its logarithm is normally distributed. The lognormal distribution is therefore applicable for positive valued quantities and has a heavy tail. σ and μ are the standard deviation and the mean of the logarithmic variable $\ln C$, respectively. The lognormal distribution occurs in variables from many fields, including biological and economic attributes (Sachs, 1984).

Figure 2 shows exemplary empirical distributions of three different types of resources to illustrate that real-world resource data can be qualitatively described by a lognormal distribution with least-squares fits revealing different σ parameters: (i) forested land area per country $\sigma = 3.83$ for the year 1991; (ii) biocapacity per country $\sigma = 1.42$ computed from the Ecological Footprint Network (Ewing et al., 2008), representing the capacity of ecosystems to regenerate what people extract; and (iii) total renewable water resources data $\sigma = 1.98$ characterizing the maximum yearly amount of water available to each country for the year 2012. Although the agreement between the lognormal distribution and the data is far from perfect, Fig. 2 supports the use of a lognormal model for resource heterogeneity in modeling our stylized social–ecological system.

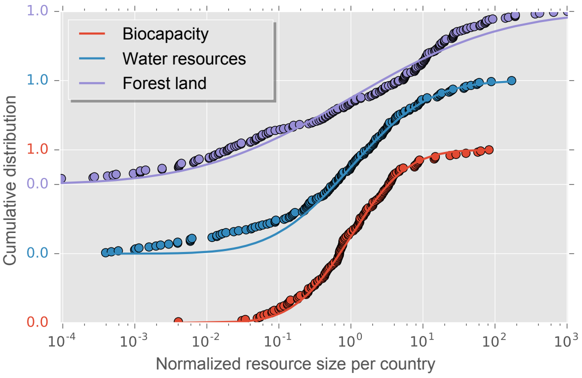


Figure 2. Empirical resource data per country normalized to the respective average (dots) together with least-squares-fitted lognormal distributions (lines): biocapacity ($\sigma = 1.42$, for the year 2007) computed from the Ecological Footprint Network (Ewing et al., 2008) represents the capacity of ecosystems to regenerate what people demand from them; total renewable water resources ($\sigma = 1.98$, for the year 2012) corresponds to the maximum theoretical yearly amount of water actually available for a country; forest land area per country ($\sigma = 3.83$, for the year 1991). The data are normalized to yield the same parameter $\mu = 0$ of the lognormal distribution and are shifted along the y axis for the sake of visibility. Note that the data qualitatively fit the lognormal distribution and that they give different values for the σ parameters of the lognormal distribution.

We utilize this distribution to investigate how resource heterogeneity affects the behavior of the model in comparison to the frequently studied homogeneous case. We systematically increase parameter σ of the lognormal distribution, which can be interpreted as a resource heterogeneity parameter, and study the resulting behavior of the model. This is done while keeping the mean of C and, consequently, the cumulative carrying capacity of all resource stocks constant – i.e., the parameter μ was adjusted according to $\mu(\sigma) = -\sigma^2/2$, resulting in a fixed value of one for the mean of C. Hence, we only ask for the effect of different resource distributions and keep the total amount of available resource stock constant.

For comparison we also present results for non-heavy-tailed resource capacities

$$C = |C^{\mathrm{tmp}}|, \quad \text{where} \quad C^{\mathrm{tmp}} \sim \mathcal{N}(C^{\mathrm{tmp}}; \mu_{\mathcal{N}}, \sigma_{\mathcal{N}}) = \frac{1}{\sigma_{\mathcal{N}}\sqrt{2\pi}} \exp\left[-\frac{(C^{\mathrm{tmp}} - \mu_{\mathcal{N}})^2}{2\sigma_{\mathcal{N}}^2}\right], \qquad (5)$$

where $\mu_{\mathcal{N}}$ now denotes the mean and $\sigma_{\mathcal{N}}$ the standard deviation of the underlying normal distribution. We also keep the mean fixed ($\mu_{\mathcal{N}} = 1$) and systematically increase the resource heterogeneity $\sigma_{\mathcal{N}}$ on comparable ranges of variances for both – normal and lognormal – distributions. Since the

Reproduced from: W. Barfuss et al., Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dyn., vol. 8, no. 2, pp. 255–264, 2017, doi: 10.5194/esd-8-255-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

normal distribution is not bounded to positive values, we use the absolute value of the drawn random variable as the resource's carrying capacity C.

2.3 Model parameterization and simulation protocol

A model run starts with an initial condition of stocks $s_i(0)$ uniformly distributed between 0 and C_i and harvesting strategies $S_i(0)$ drawn with a probability of 0.5 for a sustainable strategy $S_i = 1$ or a non-sustainable strategy $S_i = 0$. From the initial conditions, the model will converge to the steady state at t_f, where no further updates of strategy can occur. This is the case because the social network will consist solely of disconnected components with only one harvesting strategy (including the case of one single component) (Wiedermann et al., 2015). The remaining model parameters are the number of nodes $N = 500$, mean degree $\bar{k} = 20$, imitation tendency $\gamma = 1$, and ecological growth rate $g_i = 1$ for $i = 1, \ldots N$, which are kept fixed throughout the analysis. To account for the stochasticity inherent in the model, we perform $R = 250$ runs for each parameter setting of interest. We are interested in the fraction of sustainable harvesting nodes at the steady state,

$$\langle S(t_f)\rangle_{N,R} = \left\langle \frac{1}{N}\sum_{i=1}^{N} S_i(t_f)\right\rangle_R , \qquad (6)$$

averaged over all ensemble runs R. $\langle S(t_f)\rangle_{N,R}$ is bounded between one and zero, where $\langle S(t_f)\rangle_{N,R} = 1(0)$ denotes a completely (non-)sustainable regime.

3 Results and discussion

3.1 Social interaction timescale–homophily parameter space

First, we study how the fraction of sustainable harvesting nodes at the steady state $\langle S(t_f)\rangle_{N,R}$ (Eq. 6) behaves in the parameter subspace spanned by the rewiring probability ϕ (as a measure of the degree of homophily) and the average social interaction timescale τ for vanishing resource heterogeneity ($\sigma = 0.01$) (Fig. 3a).

Four qualitatively different regimes can be observed: the sustainable regime in blue, the non-sustainable or collapse regime in red, and the transition regime in white between these, as well as, for sufficiently large ϕ, the network fragmentation regime. The latter occurs since for large ϕ, social dynamics are dominated by homophily and, hence, by the process of social network rewiring, and thus negligibly few changes in strategy occur. The steady state is reached by a fragmentation of the network into at least one purely sustainable and at least one purely non-sustainable component of comparable size.

In turn, for smaller ϕ the effect of homophily is sufficiently weak such that most agents remain connected to a single component in the social network. The steady state is reached with a large connected network component. Here, large interaction timescales τ lead to a sustainable regime. This is because the comparisons of harvest rates typically happen when the logistic resource has been harvested for a sufficiently long time to reveal that the harvest rate converges to a positive value for a sustainable strategy, whereas for a non-sustainable strategy it converges to zero.

Our main focus lies on the emergent properties of our model from a complex system's perspective. Hence, we do not claim that any quantitative choice of parameters is based on real-world assumptions. Rather, we focus here on qualitative observations in terms of general parameter regimes which in correspondence with the arbitrarily chosen ecological timescale cause a certain differential outcome of the model. However, in order to qualitatively compare our model with some real-world observations, we first look at the timescale of social updates τ. It has been suggested than modern lifestyles are dominated by a social acceleration (Rosa, 2013). Simultaneously, the pressure humanity is putting on the planet (Steffen et al., 2004) has experienced a great acceleration (Steffen et al., 2015a). This can be interpreted such that faster social timescales τ lead to a non-sustainable regime, as observed in our model (see Fig. 3). Viewed with caution, the mechanisms in our model might be a possible explanation of this phenomenon. In any case, it highlights the importance of well-interacting social timescales with ecological ones. Since ecological timescales (e.g., the seasonal cycle) are difficult to influence, this suggests to take social timescales (e.g., election cycles, fashion trends, product launches) into account for possible policy interventions. Therefore, it might be worthwhile to study the relationship between social and ecological timescales more intensively to identify suitable policy actions for the benefit of a sustainable system.

We furthermore observe a linear relationship between critical parameters ϕ_c and τ_c where the transition between collapse and sustainable regimes occurs (Fig. 3). This result can be explained by the rate at which strategy changes happen. For $\phi = 0$, the transition occurs at $1/\tau \approx 1$, i.e., the ecological growth rate. For $\phi > 0$, imitation interactions happen at a rate $(1-\phi)/\tau$ (Wiedermann et al., 2015) since the network rewires with probability ϕ and, hence, imitation takes place with probability $1-\phi$. Hence, the effective imitation rate $(1-\phi)/\tau$ equals approximately 1 (the ecological growth rate) in the transition regime, which explains the linear dependence between the two social parameters.

In other words, the homophily process in our model is beneficial for reaching the sustainable regime, where all agents harvest their resource gaining the maximum sustainable yield. All stochasticity and inherent shocks towards this sustainable steady state are absorbed and not affecting the final outcome. In this sense the sustainable regime can be described as resilient. This aligns with the findings of Newig et al. (2010), who (although from a different perspective)

Reproduced from: W. Barfuss et al., Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dyn., vol. 8, no. 2, pp. 255–264, 2017, doi: 10.5194/esd-8-255-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

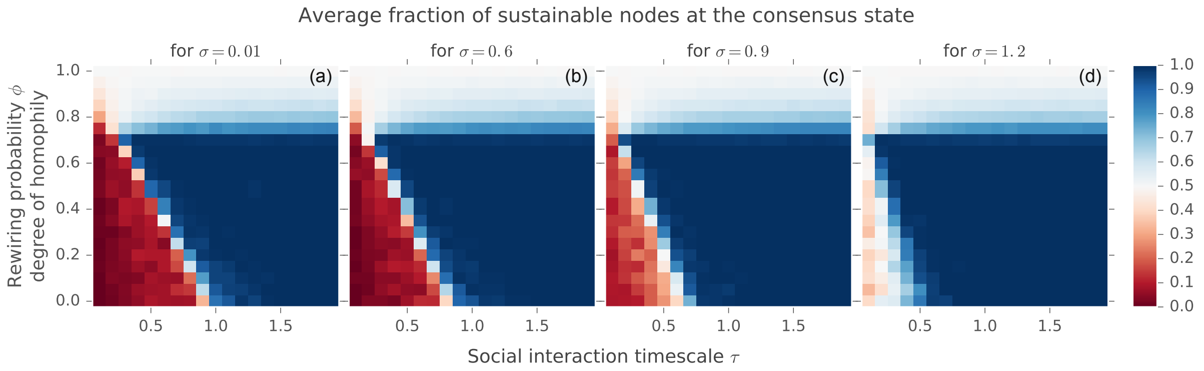


Figure 3. Social interaction timescale–homophily parameter space. Average fraction of sustainable harvesting agents in the steady state depending on the social network rewiring probability ϕ (measuring the degree of homophily) and the social interaction timescale τ for four distinct levels of resource heterogeneity (**a**: $\sigma = 0.01$; **b**: $\sigma = 0.6$; **c**: $\sigma = 0.9$; **d**: $\sigma = 1.2$). One observes four qualitatively different regimes: (i) the sustainable regime for $\phi \lesssim 0.8$ and sufficiently large (slow) τ in blue, (ii) the non-sustainable or collapse regime for $\phi \lesssim 0.8$ and sufficiently small (fast) τ in red, and (iii) in between both the transition regime in white and (iv) the network fragmentation regime for $\phi \gtrsim 0.8$, also in white.

hypothesize that homophily has a beneficial effect on the resilience of a social–ecological network. Furthermore, one can interpret a large homophily parameter ϕ as the agents' means to protect themselves against the fast and free exchange of harvesting strategies. Along similar lines, it has been found that individuals with more environmental concerns also hold more protectionist policy preferences (Bechtel et al., 2012). Our model suggests one possible mechanism for how these relationships might come into place. However, it needs to be stated that too high a rewiring probability leads to a fragmentation of the social network into smaller groups of disjoint strategies, preventing the opportunity of a completely sustainable outcome. Thus, network adaptation at very high rates should be avoided for the sake of knowledge exchange and consensus formation.

Overall, these results demonstrate that immaterial processes distinct from macroeconomic optimization paradigms and residing exclusively in the social sphere, such as homophily and imitation, are capable of determining the eventual state of a material renewable resource. Thereby, these processes are able to govern a coupled social–ecological system such that full sustainability and total collapse are possible outcomes within the investigated social parameter space. Additionally, they show how the interaction of different social processes such as strategy imitation and homophily is able to shape the sustainable regime. This suggests that socio-cultural processes should be considered as a potentially important part of feedback loops also in more elaborate models of the "World–Earth" system.

3.2 Systematic analysis of resource heterogeneity

We next investigate how the transition between sustainable and non-sustainable steady states depends on the parameter σ governing resource heterogeneity. We observe a qualitatively similar structure of parameter space for varying degrees of resource heterogeneity, but observe a decreasing extent of the non-sustainable regime for increasing σ (Fig. 3a–d).

A more systematic analysis examines the average fraction of sustainable harvesting nodes at the consensus state $\langle S(t_f)\rangle_{N,R}$ for several segments of the parameter space spanned by τ, ϕ and the resource heterogeneity parameters σ ($\sigma_{\mathcal{N}}$) – i.e., results are shown for both lognormally and normally distributed resource carrying capacities (Fig. 4). The ranges of σ for the lognormal and $\sigma_{\mathcal{N}}$ for the normal distribution are chosen such that they correspond to comparable standard deviations.

This analysis allows for explicitly showing the effect of resource heterogeneity on the critical values τ_c (Fig. 4a, c) and ϕ_c (Fig. 4b, d), where the transition from the non-sustainable to the sustainable regime occurs. In general, the larger the σ ($\sigma_{\mathcal{N}}$), the smaller the τ_c and ϕ_c. In other words, a sustainable steady state can be achieved for faster social interactions and smaller degrees of homophily the larger the resource heterogeneity is. The critical effective update timescale $\tau/(1-\phi) \overset{!}{=} \tau_{\text{eff,crit}}$ decreases to faster update times. This behavior is more pronounced for the lognormal distribution (Fig. 4a, b) than for the normal one (Fig. 4c, d) and can be explained by the heavy tails of the lognormal distribution. For a sufficiently large resource heterogeneity σ there is a sufficiently high probability that some agents will be assigned a comparably large resource capacity. Non-sustainable harvesting agents exploit their resources exponentially fast in time, whereas sustainable harvesting agents with comparably large resource capacity can retain their resource stock at a level that is still sufficiently large to convince other agents to become sustainable as well.

Reproduced from: W. Barfuss et al., Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dyn., vol. 8, no. 2, pp. 255–264, 2017, doi: 10.5194/esd-8-255-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

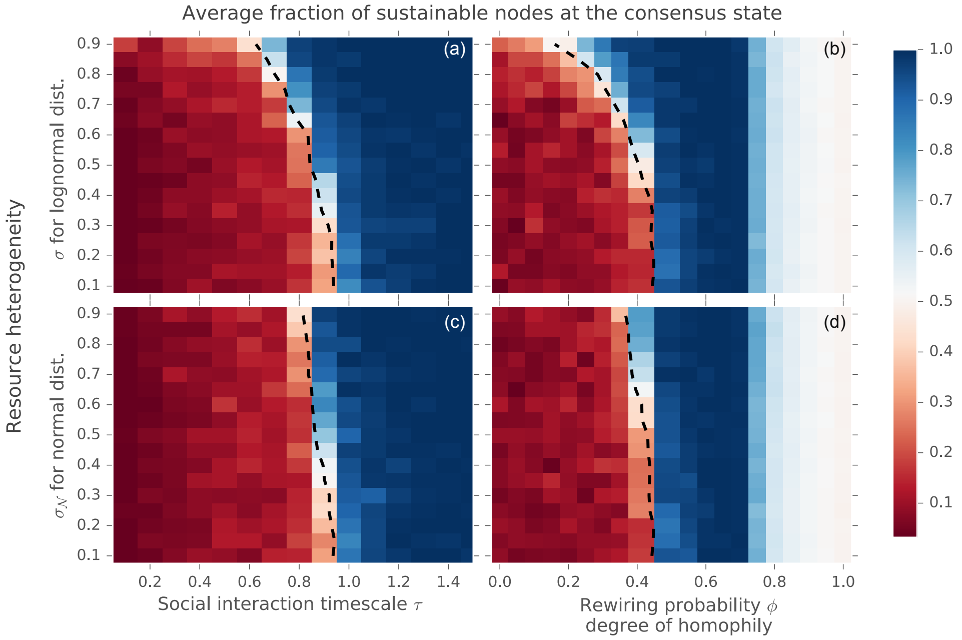


Figure 4. Effects of resource heterogeneity. Average fraction of sustainable harvesting nodes at the steady state for several segments of parameter space: **(a, b)** for (heavy-tailed) lognormally distributed capacities and **(c, d)** for (non-heavy-tailed) normally distributed capacities. Parameter spaces spanned by **(a, c)** social interaction timescale τ and resource heterogeneity σ ($\sigma_{\mathcal{N}}$) for rewiring probability $\phi = 0$, and **(b, d)** by ϕ and σ ($\sigma_{\mathcal{N}}$) for $\tau = 0.5$. The ranges of σ and $\sigma_{\mathcal{N}}$ were chosen such that the standard deviations of both distributions are comparable. For both distributions, the mean was fixed to 1. The dashed black lines indicate the linearly interpolated 50 % average fraction of sustainable nodes. Note the considerable effect the lognormal resource capacity distribution (in comparison to the normal distribution) has on the critical values of τ and ϕ, where the transition between the sustainable and the non-sustainable regime occurs.

At first, the observation that heterogeneity in access to private resources is enlarging the sustainable regime might be contradictory to reasonable assumptions. However, it demonstrates the value of a thorough system's analysis and being critical about one's own perception of what is reasonable. Cautiously comparing this phenomenon with the real world one can interpret the size of the resource capacity as the effective economic power of international macro-agents, such as world regions or nation states. This is justified, since we do not model any other economic processes but resource extraction – for example, trade, innovation and labor. The agents with comparably large economic power that employ a sustainable strategy have greater persuasive power than sustainable agents with smaller economic power. The German energy transition and its perceived impact on other countries regarding the transition towards a sustainable energy supply might be a real-world example where a country that is comparably strong economically also exerts comparably large persuasive power over other countries to move forward towards sustainable energy supply.

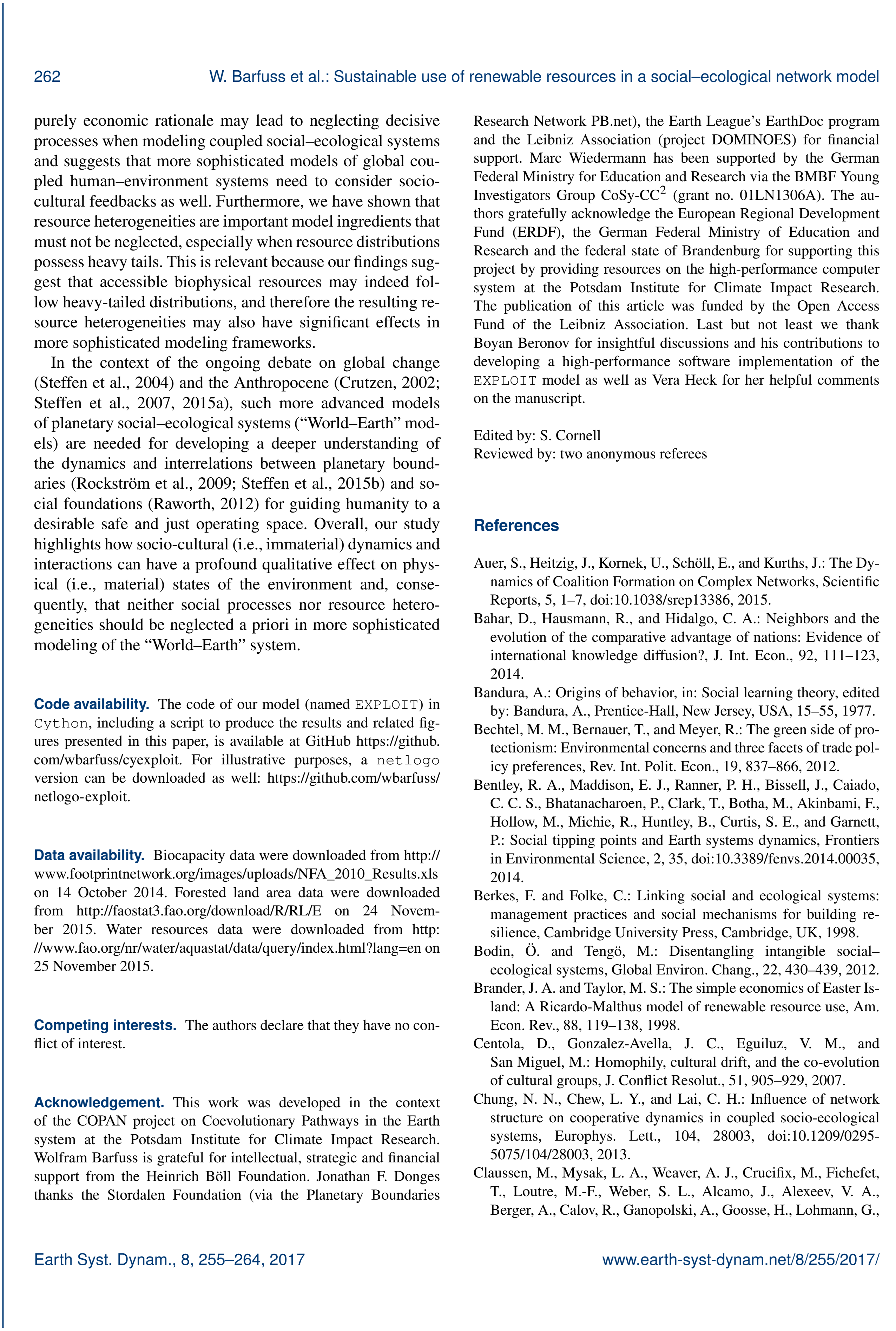
Overall, heterogeneity to resource access in our model demonstrates how comparably few sustainable first movers with a large resource capacity are also able to shift the overall system toward a sustainable state at fast social interaction rates.

4 Conclusions

In this paper, we have studied how social–ecological thresholds between sustainable and non-sustainable resource-use regimes depend on networked social interactions (related to imitation of harvesting strategies and homophily) under conditions of resource heterogeneity. We have employed a stylized model of networked agents harvesting private renewable resources with either a sustainable or non-sustainable strategy. The strategies employed by the agents are updated through a social learning process on an adaptive social network reflecting an interconnected society. Resource heterogeneity is operationalized by lognormally and normally distributed carrying capacities of the resources.

We have shown that the properties of social processes such as strategy formation by bounded rational imitation and homophilic social network adaptation alone can precondition the long-term state of renewable resources with outcomes ranging from environmental collapse to sustainability. This observation is important because it shows that following a

Reproduced from: W. Barfuss et al., Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dyn., vol. 8, no. 2, pp. 255–264, 2017, doi: 10.5194/esd-8-255-2017. Published under Creative Commons Attribution License 3.0 (CC BY).



purely economic rationale may lead to neglecting decisive processes when modeling coupled social–ecological systems and suggests that more sophisticated models of global coupled human–environment systems need to consider socio-cultural feedbacks as well. Furthermore, we have shown that resource heterogeneities are important model ingredients that must not be neglected, especially when resource distributions possess heavy tails. This is relevant because our findings suggest that accessible biophysical resources may indeed follow heavy-tailed distributions, and therefore the resulting resource heterogeneities may also have significant effects in more sophisticated modeling frameworks.

In the context of the ongoing debate on global change (Steffen et al., 2004) and the Anthropocene (Crutzen, 2002; Steffen et al., 2007, 2015a), such more advanced models of planetary social–ecological systems ("World–Earth" models) are needed for developing a deeper understanding of the dynamics and interrelations between planetary boundaries (Rockström et al., 2009; Steffen et al., 2015b) and social foundations (Raworth, 2012) for guiding humanity to a desirable safe and just operating space. Overall, our study highlights how socio-cultural (i.e., immaterial) dynamics and interactions can have a profound qualitative effect on physical (i.e., material) states of the environment and, consequently, that neither social processes nor resource heterogeneities should be neglected a priori in more sophisticated modeling of the "World–Earth" system.

Code availability. The code of our model (named `EXPLOIT`) in `Cython`, including a script to produce the results and related figures presented in this paper, is available at GitHub https://github.com/wbarfuss/cyexploit. For illustrative purposes, a `netlogo` version can be downloaded as well: https://github.com/wbarfuss/netlogo-exploit.

Data availability. Biocapacity data were downloaded from http://www.footprintnetwork.org/images/uploads/NFA_2010_Results.xls on 14 October 2014. Forested land area data were downloaded from http://faostat3.fao.org/download/R/RL/E on 24 November 2015. Water resources data were downloaded from http://www.fao.org/nr/water/aquastat/data/query/index.html?lang=en on 25 November 2015.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgement. This work was developed in the context of the COPAN project on Coevolutionary Pathways in the Earth system at the Potsdam Institute for Climate Impact Research. Wolfram Barfuss is grateful for intellectual, strategic and financial support from the Heinrich Böll Foundation. Jonathan F. Donges thanks the Stordalen Foundation (via the Planetary Boundaries Research Network PB.net), the Earth League's EarthDoc program and the Leibniz Association (project DOMINOES) for financial support. Marc Wiedermann has been supported by the German Federal Ministry for Education and Research via the BMBF Young Investigators Group CoSy-CC^2 (grant no. 01LN1306A). The authors gratefully acknowledge the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the federal state of Brandenburg for supporting this project by providing resources on the high-performance computer system at the Potsdam Institute for Climate Impact Research. The publication of this article was funded by the Open Access Fund of the Leibniz Association. Last but not least we thank Boyan Beronov for insightful discussions and his contributions to developing a high-performance software implementation of the `EXPLOIT` model as well as Vera Heck for her helpful comments on the manuscript.

Edited by: S. Cornell
Reviewed by: two anonymous referees

References

Auer, S., Heitzig, J., Kornek, U., Schöll, E., and Kurths, J.: The Dynamics of Coalition Formation on Complex Networks, Scientific Reports, 5, 1–7, doi:10.1038/srep13386, 2015.

Bahar, D., Hausmann, R., and Hidalgo, C. A.: Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion?, J. Int. Econ., 92, 111–123, 2014.

Bandura, A.: Origins of behavior, in: Social learning theory, edited by: Bandura, A., Prentice-Hall, New Jersey, USA, 15–55, 1977.

Bechtel, M. M., Bernauer, T., and Meyer, R.: The green side of protectionism: Environmental concerns and three facets of trade policy preferences, Rev. Int. Polit. Econ., 19, 837–866, 2012.

Bentley, R. A., Maddison, E. J., Ranner, P. H., Bissell, J., Caiado, C. C. S., Bhatanacharoen, P., Clark, T., Botha, M., Akinbami, F., Hollow, M., Michie, R., Huntley, B., Curtis, S. E., and Garnett, P.: Social tipping points and Earth systems dynamics, Frontiers in Environmental Science, 2, 35, doi:10.3389/fenvs.2014.00035, 2014.

Berkes, F. and Folke, C.: Linking social and ecological systems: management practices and social mechanisms for building resilience, Cambridge University Press, Cambridge, UK, 1998.

Bodin, Ö. and Tengö, M.: Disentangling intangible social–ecological systems, Global Environ. Chang., 22, 430–439, 2012.

Brander, J. A. and Taylor, M. S.: The simple economics of Easter Island: A Ricardo-Malthus model of renewable resource use, Am. Econ. Rev., 88, 119–138, 1998.

Centola, D., Gonzalez-Avella, J. C., Eguiluz, V. M., and San Miguel, M.: Homophily, cultural drift, and the co-evolution of cultural groups, J. Conflict Resolut., 51, 905–929, 2007.

Chung, N. N., Chew, L. Y., and Lai, C. H.: Influence of network structure on cooperative dynamics in coupled socio-ecological systems, Europhys. Lett., 104, 28003, doi:10.1209/0295-5075/104/28003, 2013.

Claussen, M., Mysak, L. A., Weaver, A. J., Crucifix, M., Fichefet, T., Loutre, M.-F., Weber, S. L., Alcamo, J., Alexeev, V. A., Berger, A., Calov, R., Ganopolski, A., Goosse, H., Lohmann, G.,

Reproduced from: W. Barfuss et al., Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dyn., vol. 8, no. 2, pp. 255–264, 2017, doi: 10.5194/esd-8-255-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

Lunkeit, F., Mokhov, I. I., Petoukhov, V., Stone, P., and Wang, Z.: Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models, Clim. Dynam., 18, 579–586, 2002.

Crutzen, P. J.: Geology of mankind, Nature, 415, 23–23, 2002.

Erdös, P. and Rényi, A.: On the evolution of random graphs, Publ. Math. Inst. Hungar. Acad. Sci, 5, 17–61, 1960.

Ewing, B., Goldfinger, S., Wackernagel, M., Stechbart, M., Rizk, S. M., Reed, A., and Kitzes, J.: The Ecological Footprint Atlas 2008, Global Footprint Network, Oakland, 2008.

Gross, T. and Blasius, B.: Adaptive coevolutionary networks: a review, J. R. Soc. Interface, 5, 259–271, 2008.

Hardin, G.: The tragedy of the commons. The population problem has no technical solution; it requires a fundamental extension in morality, Science, New York, NY, 162, 1243, 1968.

Holme, P. and Newman, M. E.: Nonequilibrium phase transition in the coevolution of networks and opinions, Phys. Rev. E, 74, 056108, doi:10.1103/PhysRevE.74.056108, 2006.

Janssen, M. A., Bodin, Ö., Anderies, J. M., Elmqvist, T., Ernstson, H., McAllister, R. R., Olsson, P., and Ryan, P.: Toward a network perspective of the study of resilience in social-ecological systems, Ecol. Soc., 11, 15, http://www.ecologyandsociety.org/vol11/iss1/art15/, 2006.

Keeling, M. J.: Multiplicative moments and measures of persistence in ecology, J. Theor. Biol., 205, 269–281, 2000.

Lade, S. J., Tavoni, A., Levin, S. A., and Schlüter, M.: Regime shifts in a social-ecological system, Theor. Ecol., 6, 359–372, 2013.

Lade, S. J., Örjan, B., Donges, J. F., Enfors, E., Galafassi, D., Olsson, P., Österblom, H., and Schlüter, M.: Modelling social-ecological transformations: an adaptive network proposal, in review, 2017.

Malthus, T. R.: An essay on the principle of population, as it affects the future improvement of society. With remarks on the speculations of Mr. Godwin, edited by: Condorcet, M. and other writers, J. Johnson, London, 1798.

McPherson, M., Smith-Lovin, L., and Cook, J. M.: Birds of a feather: Homophily in social networks, Annu. Rev. Sociol., 27, 415–444, 2001.

Meadows, D. H., Goldsmith, E. I., and Meadows, P.: The limits to growth, vol. 381, Earth Island Limited, London, 1972.

Newig, J., Günther, D., and Pahl-Wostl, C.: Synapses in the network: learning in governance networks in the context of environmental management, Ecol. Soc., 15, 24, http://www.ecologyandsociety.org/vol15/iss4/art24/, 2010.

Ostrom, E.: Governing the commons, Cambridge University Press, Cambridge, UK, 2015.

Perman, R., Ma, Y., McGilvray, J., and Common, M.: Natural resource and environmental economics, Pearson Education, Harlow, UK, 2003.

Raworth, K.: A safe and just space for humanity: can we live within the doughnut, Oxfam Policy and Practice: Climate Change and Resilience, 8, 1–16, 2012.

Rockström, J., Steffen, W., Noone, K., Persson, A., Chapin, F. S., Lambin, E. F., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sorlin, S., Snyder, P. K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R. W., Fabry, V. J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., and Foley, J. A.: A safe operating space for humanity, Nature, 461, 472–475, 2009.

Rosa, H.: Social acceleration: A new theory of modernity, Columbia University Press, New York, 2013.

Sachs, L.: Applied Statistics, Springer, New York, Berlin, Heidelberg, Tokyo, 1984.

Sayama, H., Pestov, I., Schmidt, J., Bush, B. J., Wong, C., Yamanoi, J., and Gross, T.: Modeling complex systems with adaptive networks, Comput. Math. Appl., 65, 1645–1664, 2013.

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., and Walker, B.: Catastrophic shifts in ecosystems, Nature, 413, 591–596, 2001.

Schellnhuber, H.-J.: Discourse: Earth System analysis – The scope of the challenge, in: Earth System Analysis, edited by: Schellnhuber, H.-J. and Wenzel, V., Springer, Berlin, 3–195, 1998.

Schellnhuber, H.-J.: Earth system analysis and the second Copernican revolution, Nature, 402, C19–C23, 1999.

Schellnhuber, H. J.: Tipping elements in the Earth System, P. Natl. Acad. Sci. USA, 106, 20561–20563, 2009.

Schleussner, C.-F., Donges, J. F., Engemann, D. A., and Levermann, A.: Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure, Scientific Reports, 6, 30790, doi:10.1038/srep30790, 2016.

Snijders, T. A., Van de Bunt, G. G., and Steglich, C. E.: Introduction to stochastic actor-based models for network dynamics, Soc. Networks, 32, 44–60, 2010.

Steffen, W., Sanderson, A., Tyson, P. D., Jäger, J., Matson, P. A., Moore III, B., Oldfield, F., Richardson, K., Schellnhuber, H. J., Turner II, B. L., and Wasson, R. J.: Global Change and the Earth System: A Planet Under Pressure, Springer, Berlin, 2004.

Steffen, W., Crutzen, P. J., and McNeill, J. R.: The Anthropocene: are humans now overwhelming the great forces of nature, Ambio, 36, 614–621, 2007.

Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O., and Ludwig, C.: The trajectory of the Anthropocene: The Great Acceleration, The Anthropocene Review, 2, 81–98, 2015a.

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., de Vries, W., de Wit, C. A., et al.: Planetary boundaries: Guiding human development on a changing planet, Science, 347, 1259855, doi:10.1126/science.1259855, 2015b.

Sugiarto, H. S., Chung, N. N., Lai, C. H., and Chew, L. Y.: Socioecological regime shifts in the setting of complex social interactions, Phys. Rev. E, 91, 062804, doi:10.1103/PhysRevE.91.062804, 2015.

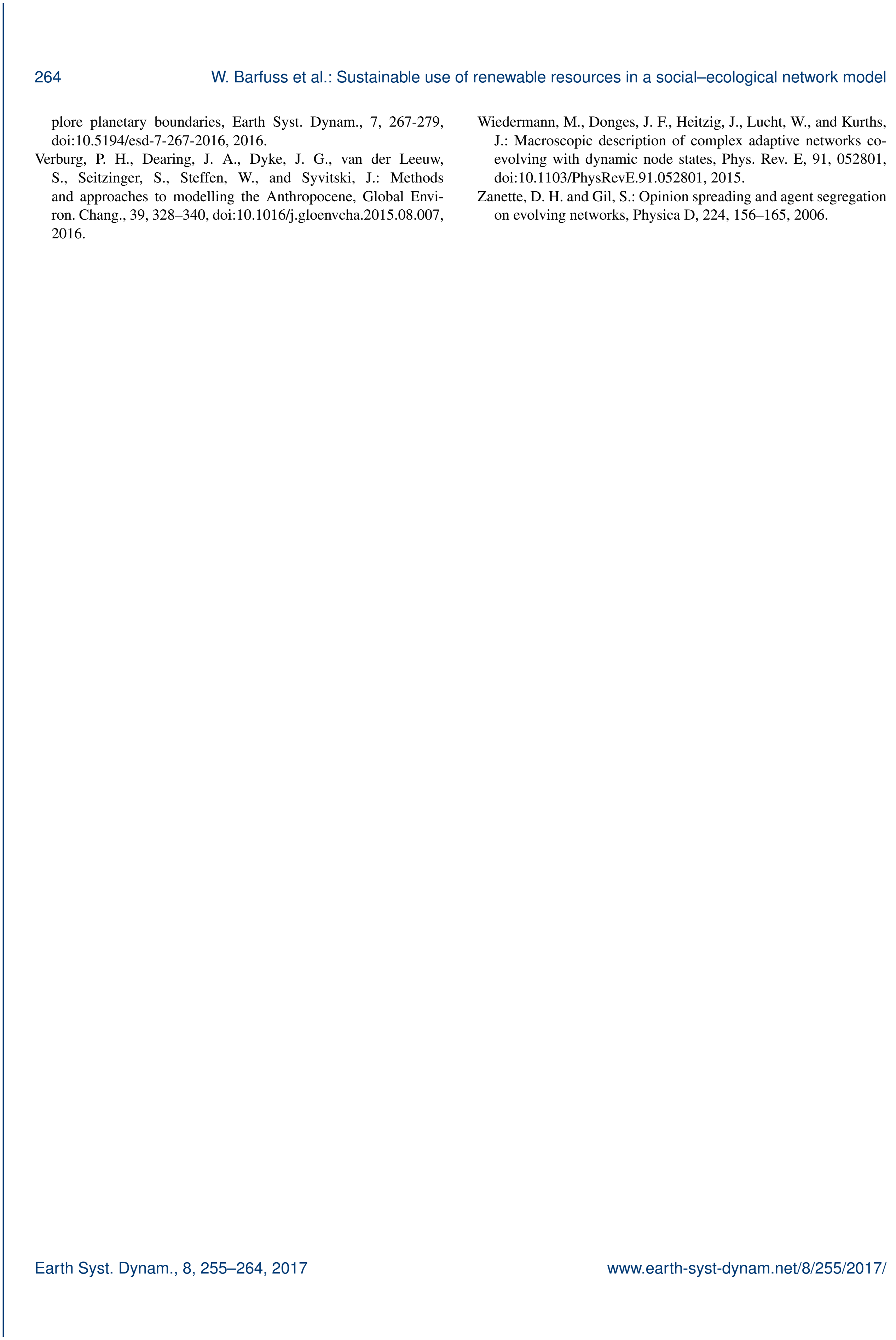
Tavoni, A., Schlüter, M., and Levin, S.: The survival of the conformist: social pressure and renewable resource management, J. Theor. Biol., 299, 152–161, 2012.

Traulsen, A., Semmann, D., Sommerfeld, R. D., Krambeck, H.-J., and Milinski, M.: Human strategy updating in evolutionary games, P. Natl. Acad. Sci. USA, 107, 2962–2966, 2010.

van Vuuren, D. P., Bayer, L. B., Chuwah, C., Ganzeveld, L., Hazeleger, W., van den Hurk, B., van Noije, T., O'Neill, B., and Strengers, B. J.: A comprehensive view on climate change: coupling of earth system and integrated assessment models, Environ. Res. Lett., 7, 024012, doi:10.1088/1748-9326/7/2/024012, 2012.

van Vuuren, D. P., Lucas, P. L., Häyhä, T., Cornell, S. E., and Stafford-Smith, M.: Horses for courses: analytical tools to ex-

Reproduced from: W. Barfuss et al., Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dyn., vol. 8, no. 2, pp. 255–264, 2017, doi: 10.5194/esd-8-255-2017. Published under Creative Commons Attribution License 3.0 (CC BY).



plore planetary boundaries, Earth Syst. Dynam., 7, 267-279, doi:10.5194/esd-7-267-2016, 2016.

Verburg, P. H., Dearing, J. A., Dyke, J. G., van der Leeuw, S., Seitzinger, S., Steffen, W., and Syvitski, J.: Methods and approaches to modelling the Anthropocene, Global Environ. Chang., 39, 328–340, doi:10.1016/j.gloenvcha.2015.08.007, 2016.

Wiedermann, M., Donges, J. F., Heitzig, J., Lucht, W., and Kurths, J.: Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E, 91, 052801, doi:10.1103/PhysRevE.91.052801, 2015.

Zanette, D. H. and Gil, S.: Opinion spreading and agent segregation on evolving networks, Physica D, 224, 156–165, 2006.

Reproduced from: W. Barfuss et al., Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dyn., vol. 8, no. 2, pp. 255–264, 2017, doi: 10.5194/esd-8-255-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

IOP Publishing *Environ. Res. Lett.* **12** (2017) 074020 https://doi.org/10.1088/1748-9326/aa7581

Environmental Research Letters

OPEN ACCESS

RECEIVED
3 February 2017

REVISED
15 May 2017

ACCEPTED FOR PUBLICATION
30 May 2017

PUBLISHED
11 July 2017

LETTER

Sustainability, collapse and oscillations in a simple World-Earth model

Jan Nitzbon[1,2,3,5], Jobst Heitzig[2] and Ulrich Parlitz[1,4]

[1] Institute for Nonlinear Dynamics, Faculty of Physics, University of Göttingen, Friedrich-Hund-Platz 1, 37077 Göttingen, Germany
[2] Potsdam Institute for Climate Impact Research, P.O. Box 601203, 14412 Potsdam, Germany
[3] Alfred Wegener Institute for Polar and Marine Research, P.O. Box 600149, 14401 Potsdam, Germany
[4] Max Planck Institute for Dynamics and Self-Organization, Am Fassberg 17, 37077 Göttingen, Germany
[5] Author to whom any correspondence should be addressed.

E-mail: jan.nitzbon@awi.de

Keywords: World-Earth modeling, anthropocene, global carbon cycle, energy transformation, coevolutionary dynamics, bifurcation analysis

Abstract

The Anthropocene is characterized by close interdependencies between the natural Earth system and the global human society, posing novel challenges to model development. Here we present a conceptual model describing the long-term co-evolution of natural and socio-economic subsystems of Earth. While the climate is represented via a global carbon cycle, we use economic concepts to model socio-metabolic flows of biomass and fossil fuels between nature and society. A well-being-dependent parametrization of fertility and mortality governs human population dynamics.

Our analysis focuses on assessing possible asymptotic states of the Earth system for a qualitative understanding of its complex dynamics rather than quantitative predictions. Low dimension and simple equations enable a parameter-space analysis allowing us to identify preconditions of several asymptotic states and hence fates of humanity and planet. These include a sustainable co-evolution of nature and society, a global collapse and everlasting oscillations.

We consider different scenarios corresponding to different socio-cultural stages of human history. The necessity of accounting for the 'human factor' in Earth system models is highlighted by the finding that carbon stocks during the past centuries evolved opposing to what would 'naturally' be expected on a planet without humans. The intensity of biomass use and the contribution of ecosystem services to human well-being are found to be crucial determinants of the asymptotic state in a (pre-industrial) biomass-only scenario without capital accumulation. The capitalistic, fossil-based scenario reveals that trajectories with fundamentally different asymptotic states might still be almost indistinguishable during even a centuries-long transient phase. Given current human population levels, our study also supports the claim that besides reducing the global demand for energy, only the extensive use of renewable energies may pave the way into a sustainable future.

1. Introduction

The impacts humankind exerts on nature on a planetary scale have become so grave that an entirely new geological epoch—the Anthropocene—has been proclaimed [1], characterized by strong nature-society interrelations. Independent of whether the Anthropocene indeed depicts a novel geological epoch or not [2–5], predicting Earth's future with models necessitates recognizing the influences humans exert on it and vice versa. This qualitatively new relation between humans and nature poses a huge challenge for the development of suitable models, demanding a balanced representation of both the *natural sphere* (ecosphere, 'Earth') and the *human sphere* (anthroposphere, 'World') and a holistic system's perspective

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

[6–9]. Many models of the natural Earth system (e.g. general circulation models (GCMs) or Earth system models of intermediate complexity (EMICs)) include human impacts only as an exogenous driver, e.g. in the form of emission scenarios [10]. Integrated assessment models (IAMs) on the other hand try to simulate and/or optimize the future economic evolution under changing environmental conditions on multiple decades [11]. However, only few modelling attempts aim at a balanced representation of natural and socio-economic dynamics on centennial to millennial time scales [12–16]. Conceptual World-Earth models like the one presented here try to fill this gap in the model landscape and thereby contribute to modelling the Anthrophocene.

Complementary to the development of useful models of World-Earth *dynamics* stands the challenge to identify a *desirable condition* of the World-Earth system. The concept of Planetary Boundaries is a major advance in this direction regarding the natural dimension [17–19]. It states that during the holocene several aggregate indicators of the Earth's state stayed within certain limits which define a kind of 'safe operating space' to which humanity is adapted and which should not be transgressed. Within the framework of the 'Oxfam doughnut' these bounds are supplemented by quantitative indicators of socio-economic aspects of the world, called 'social foundations', which together are thus interpreted to define a 'safe and just operating space' [20], see also the Sustainable Development Goals [21, 22]. The state space topology and dilemmas resulting from such boundaries can be analysed if the models are not too complex [23]. Hence, while models with dozens of state variables (e.g. World3 [13], GUMBO [14]) might allow answering rather quantitative questions, they preclude analytical analyses that provide a deeper qualitative understanding of the World-Earth system. Examples for rather simple, conceptual approaches comprise the studies of local models of natural resources co-evolving with social or population dynamics [24–27], but also models which address social stratification [15] and conceptual models on a global scale [28, 29].

Our goal here is to contribute to the latter strand of literature a simple conceptual model focussing on a few globally aggregated quantities of the natural and socio-economic subsystems that appear most essential to assess the desirability of the system state in terms of population, well-being, and biosphere integrity. As well-being and biosphere integrity depend crucially on climate and natural resource use, our World-Earth model describes the temporal evolution of the global carbon cycle, human population, and the competition between the major energy sources, biomass and fossil fuels, on centennial to millennial time-scales. A particular objective of this study is to characterize the possible asymptotic paths the world could have taken, and to identify model parameters crucial for switching between these qualitatively different dynamic regimes. To be able to apply the necessary techniques from dynamical systems theory, e.g. bifurcation analysis, we keep the dimension low, using only five dynamic variables, and the equations simple.

Despite this simplicity, the model is capable of qualitatively reflecting the actual dynamics seen during different stages in human history, in particular the Holocene and the Anthropocene. For a pre-industrial society, for instance, our model saturates at a stable global population of about 200 mn, similar to the actual global population in medieval times. The model can also produce stable cycles of population growth and decline similar to the secular cycles studied by the literature reviewed in [30]. However, while that strand of research finds centennial, domestic cycles and explains them by means of socio-cultural dynamics, we rather find millenial, global cycles which are a consequence of the carbon cycle with which population dynamics interact. Thus our model can be interpreted as adding a time-delay effect to Malthusian theory, as requested in [30].

To be more precise, we combine a carbon cycle in a novel way with well-being-driven population dynamics and economic production based on energy and accumulated capital. We model the global carbon cycle similar to [31], thereby facilitating the study of carbon-related planetary boundaries [32]. While models of comparable complexity (e.g. World2 [33] or Wonderland [29]) employ rather simple parametrizations of the economic output, our approach is founded on well established concepts from economic theory. In combination with a suitable description of population dynamics we show that without an anthroposphere component the model behaviour would deviate drastically from what is observed.

The paper is structured as follows: After introducing the full model in section 2, we analyse special cases of growing complexity that roughly relate to different eras in human history in section 3 before concluding in section 4. The appendix contains details regarding the derivation of the model, the estimation of its parameters, its bifurcation analysis, and conditions for phases of superexponential growth.

2. Model

Similar to [31], our conceptual model describes the global carbon cycle via three carbon reservoirs—the terrestrial (L, plants and soils), atmospheric (A), and geological (G) carbon stocks, and describes the global population and economy via just two additional stocks, human population P and physical capital K

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

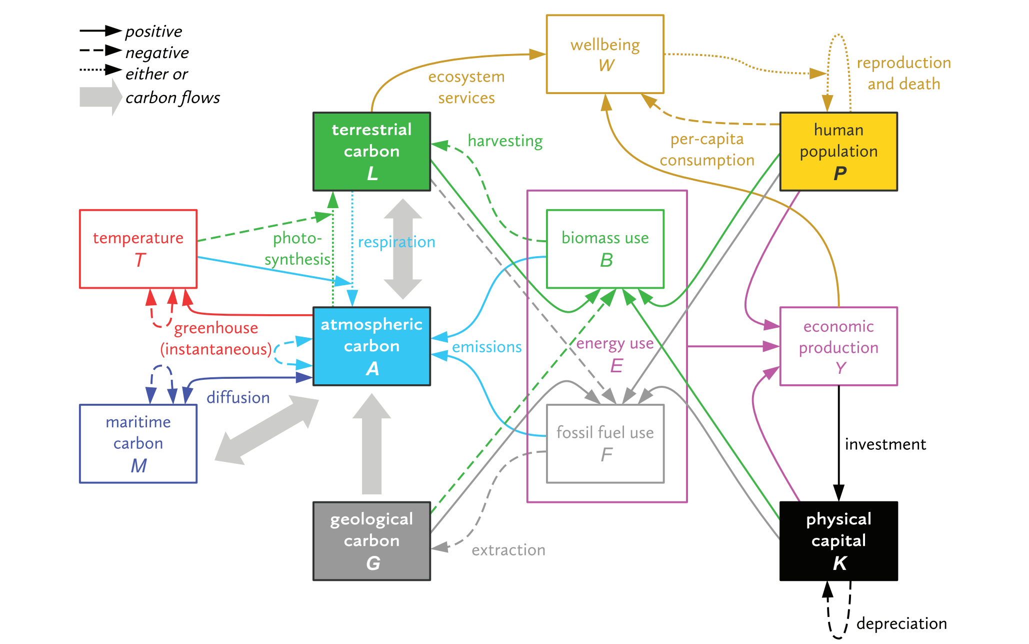


Figure 1. Overview of the model structure with five state variables (colored boxes) and several derived variables (white boxes). Arrows represent coupling processes between the variables. The left part represents the natural subsystem of the Earth (Ecosphere) via the global carbon cycle, while the right part represents socio-economic entities related to human activities in the World (Anthroposphere).

(see figure 1). Their dynamics is governed by five ordinary differential equations

$$\dot{L} = (l_0 - l_T T)\sqrt{A/\Sigma}L - (a_0 + a_T T)L - B, \quad (2.1)$$

$$\dot{A} = -\dot{L} + d(M - mA), \quad (2.2)$$

$$\dot{G} = -F, \quad (2.3)$$

$$\dot{P} = P\left(\frac{2WW_P}{W^2 + W_P^2}p - \frac{q}{W}\right), \quad (2.4)$$

$$\dot{K} = iY - kK. \quad (2.5)$$

The derived quantities of maritime carbon stock M, global mean temperature T, biomass use B, fossil fuel use F, economic production Y, and well-being W are governed by the algebraic equations

$$M = C^* - L - A - G, \quad (2.6)$$

$$T = A/\Sigma, \quad (2.7)$$

$$B = \frac{a_B}{e_B}\frac{L^2(PK)^{2/5}}{(a_B L^2 + a_F G^2)^{4/5}}, \quad (2.8)$$

$$F = \frac{a_F}{e_F}\frac{G^2(PK)^{2/5}}{(a_B L^2 + a_F G^2)^{4/5}}, \quad (2.9)$$

$$Y = y_E(e_B B + e_F F), \quad (2.10)$$

$$W = \frac{(1-i)Y}{P} + w_L\frac{L}{\Sigma}. \quad (2.11)$$

See table 1 and appendix B for parameter meanings and estimates on the basis of available real-world data. The three terms in $\dot{L}$ represent temperature-dependent photosynthesis (with atmospheric carbon fertilization) and respiration, and biomass extraction. The second term in $\dot{A}$ is diffusion at the oceans' surface. The terms in $\dot{P}$ represent well-being-dependent fertility and mortality, where fertility reaches a maximum of p at $W = W_P$ and then declines again. Finally, the terms in $\dot{K}$ are investment at a fixed savings rate and capital depreciation. Temperature T is assumed to relax instantaneously to its equilibrium value depending on A, using a nonlinear temperature scale so it is simply proportional to A. The denominator in B and F represents substitution effects in the energy sector. Economic production Y in the remaining sectors is proportional to energy input. Well-being W derives from per-capita consumption and ecosystem services assumed proportional to L. The latter comprise provisional (e.g. water, raw materials), regulating (e.g. waste decomposition) and cultural (e.g. recreational) services [34, 35]. appendix A contains a detailed motivation and derivation of the model from physical and economic principles.

3. Results

3.1. How recent centuries' carbon cycle trends oppose purely natural dynamics

We first consider the natural carbon cycle without human interference by setting $P = K = 0$. Figure 2

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

Table 1. Overview of the model parameters, their physical dimensions and the best estimate based on real-world data.

Symbol	Description	Unit (H = humans)	Estimate
Σ	available Earth surface area	km^2	1.5·10^8
C^*	total available carbon stock	GtC	5500
a_0	respiration baseline coefficient	a^{-1}	0.0298
a_T	respiration sensitivity to temperature	km^2 a^{-1} GtC^{-1}	3200
l_0	photosynthesis baseline coefficient	km a^{-1} GtC$^{-1/2}$	26.4
l_T	photosynthesis sensitivity to temperature	km^3 a^{-1} GtC$^{-3/2}$	1.1·10^6
d	diffusion rate	a^{-1}	0.01
m	solubility coefficient	1	1.5
p	fertility maximum	a^{-1}	0.04
W_P	fertility saturation well-being	\$ a^{-1} H^{-1}	2000
q	mortality baseline coefficient	\$ a^{-2} H^{-1}	20
i	investment ratio	1	0.25
k	capital depreciation rate	a^{-1}	0.1
a_B	biomass sector productivity	GJ^5 a^{-5} GtC^{-2} \$$^{-2}$ H^{-2}	varied
a_F	fossil fuel sector productivity	GJ^5 a^{-5} GtC^{-2} \$$^{-2}$ H^{-2}	varied
e_B	biomass energy density	GJ GtC^{-1}	4·10^{10}
e_F	fossil fuel energy density	GJ GtC^{-1}	4·10^{10}
y_E	economic output per energy input	\$ GJ^{-1}	147
w_L	well-being sensitivity to land carbon	\$ km^2 GtC^{-1} a^{-1} H^{-1}	varied
C^*_{PI}	total pre-industrial carbon stock	GtC	4000
b	biomass harvesting rate	GtC$^{3/5}$ a^{-1} H$^{-3/5}$	5.4·10^{-7}
y_B	economic output per biomass input	\$ GtC^{-1}	2.47·10^{11} (varied)

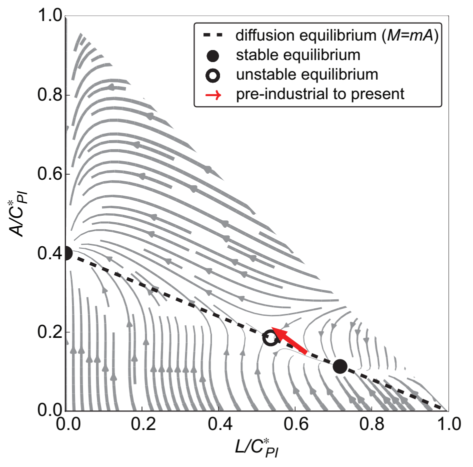


Figure 2. State space representation of the purely natural carbon cycle dynamics given by equations (2.1) and (2.2) and setting $P = K = 0$. Grey arrows show the direction of the system's evolution, thicker lines correspond to faster flow. On the black dashed line diffusion is in equilibrium. There are three equilibria of which the 'desert' state at $L_D = 0$ and the 'forest' state at $L_F \approx 0.72C^*_{PI}$ are stable. The red arrow reflects the actual evolution of the carbon pools from pre-industrial times until today. It opposes the natural direction of the flow, indicating the necessity of incorporating human activities into Earth system models. The upper right corner is not part of the state space due to the mass constraint $L + A \leq C^*_{PI}$. Parameters are set to the default values given in table 1.

shows the state space of the remaining two-dimensional system given by terrestrial (L) and atmospheric (A) carbon stocks. As $\dot{G} = 0$, the geological carbon stock G is ignored and L and A are normalized by the pre-industrial carbon amount of the (short-term) carbon cycle, C^*_{PI}.

Equilibrium states of the system require $\dot{L} = \dot{A} = 0$ so that, according to equation (2.2), net diffusion between the atmosphere and the upper ocean vanishes ($M = mA$). Solving (2.1) using the parameter values from table 1 gives three equilibria: (i) a stable *desert state* located at $L^*_{\mathrm{D}} = 0$, (ii) an intermediate unstable equilibrium at $L^*_{\mathrm{I}} \approx 0.54\ C^*_{\mathrm{PI}}$, and (iii) a stable *forest state* at $L^*_{\mathrm{F}} \approx 0.72\ C^*_{\mathrm{PI}}$. Hence, our carbon cycle component features *bistability* between a desirable (forest) and an undesirable (desert) state, to one of which the system will converge, depending on initial conditions.

The forest equilibrium represents the Holocene carbon cycle until pre-industrial times, neglecting changes in external solar forcing. During this period the exchange of carbon between the terrestrial, maritime, and atmospheric reservoirs were roughly in balance [36]. The temporal permanence during the Holocene is reflected in the model by the forest equilibrium's stability. The model will return to the forest state after *small* perturbations which might for instance occur via Volcanic eruptions or other (small) external forcing.

In contrast, the affection of the carbon cycle through human activities like land use (change) and GHG emissions constitutes a *large* perturbation of its natural dynamics. To illustrate this, the red arrow depicted in figure 2 points from the pre-industrial to the current state, far from the forest state and already in the basin of attraction of the desert state.

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

Hence, this simplistic model suggests that the carbon cycle might already be in a regime where it would collapse in the *future* even without further human influence. On the other hand, the model does not reproduce well the actual *past* evolution of the carbon cycle since the advent of the industrialization, which clearly opposes the shown 'natural' direction of the flow. For a more reliable analysis, it is thus necessary to explicitly include the human factor into our model, as demanded by [6].[6]

3.2. How oscillations may emerge in a non-fossil, pre-capitalistic global society

We thus add a dynamic human population P, interfering with the biosphere. Its only energy source is biomass, no fossil fuels ($a_F = 0$) are used yet. The global society in this scenario is assumed not to accumulate physical capital but to operate with a constant amount of capital per capita ($K \propto P$). Introducing the new parameters b and y_B, the expressions for B (2.8) and Y (2.10) read

$$B_{\mathrm{PI}} = bL^{\frac{2}{5}}P^{\frac{3}{5}}, \tag{3.1}$$

$$Y_{\mathrm{PI}} = y_B B_{\mathrm{PI}}. \tag{3.2}$$

In order to reduce the dimension of the model system without altering the qualitative (asymptotic) behaviour, the diffusion equilibrium is assumed to establish instantaneously ($d \to \infty$), implying fixed relations between the carbon stocks, $A = (C^*_{\mathrm{PI}} - L)/(1+m)$ and $M = mA$. We thus get a two-dimensional system with just L and P as dynamical variables.

In this pre-industrial scenario one can ask what will ultimately happen to a global society which solely harvests biomass. The answer strongly depends on the choice of the parameters. Consider an initial situation with $P_0 = 500\,000$ on a forested planet ($L_0 = 0.72C^*_{\mathrm{PI}} = 2880$ GtC); furthermore all parameters are set to the default values (see table 1) and ecosystem services are neglected ($w_L = 0$) (figure 3, upper right panel). Due to the abundance of resources, the population initially prospers and grows (exponentially) fast. Biomass use also increases but slower than population (equation (3.1)), so that well-being decreases as a consequence (equation 2.11); this in turn lets the population growth rate decrease. After about 600 years a maximum population of about one billion humans is reached while the terrestrial carbon stock is considerably lower than initially. Despite the following decrease in population, the pressure on the ecosphere by humans pushes the carbon cycle into the basin of attraction of the (undesirable) desert state and an unpopulated planet prevails after about 1200 years. When regarding the state space of the system (figure 3, upper left panel) it becomes clear why this *collapse* was inevitable. There simply is no *coexistence equilibrium* with $L > 0$ and $P > 0$, and even the two unpopulated forest equilibria with $L > 0$ and $P = 0$ are unstable (one in the L-, the other in the P-direction) so that only the desert state equilibrium at $L = P = 0$ is an attractor. Hence independent of the initial conditions the system will ultimately evolve to the desert state.

While such collapse has been observed historically for *local* agricultural civilizations [24], a global collapse of the terrestrial ecosystems did not occur so far. For slightly altered parameter values, an evolution of the model system occurs which matches the historic one better, until the onset of the industrialization. However, if the value of y_B (whose estimate has a high uncertainty) is halved, a *sustained coexistence* between the terrestrial ecosystems and the human population becomes possible (figure 3 middle panels). In addition to the three equilibria at the P-axis ($P = 0$), there exist two equilibria with $L > 0$ and $P > 0$ of which one is stable. Starting from the same initial state as above the system initially behaves similar, but the population rise is less extreme and humans exert less pressure on the terrestrial carbon stock. After about 400 years an equilibrium with constant carbon stocks, population and well-being is reached. The asymptotic population of about 200 mn compares nicely with actual estimates of the global population in medieval times [37], for which the non-fossil, pre-capitalistic model scenario seems adequate. A long period of stagnating socio-economic observables is also in line with the Malthusian population model [38].

Like Malthus, we identified well-being (which determines fertility and mortality, see (2.4)) with per-capita consumption so far. It is, however, reasonable to assume that the integrity of nature also contributes to human well-being via ecosystem services (e.g. the provision of forage to hunter-gatherer communities). Hence we consider a third setting in which well-being is dominated by ecosystem services by choosing $w_L > 0$ and a low value for y_B (figure 3, lower panels). The phase portrait qualitatively differs from both previous cases as it features an attracting *limit-cycle* but no stable coexistence equilibrium. Hence there are trajectories—such as the shown one—which are characterized by sustained *oscillations* of all variables. As before, population rises until it reaches a maximum of about 500 mn humans after about 1500 years. The growing biomass consumption is accompanied by decreasing well-being and—with a short delay—decreasing population. P declines until it reaches a minimum after another approximately 800 years, now taking pressure from the terrestrial carbon stock, which is thus able to recover. This in turn directly increases well-being via the contribution of ecosystem services, allowing population to recover as well. These feedbacks lead to oscillations with a period of about

[6] Note that the subsequent analyses focus on the parametrization of the socio-economic model components while the in-depth study and advancement of its natural component (e.g. representation of the global water cycle) is not within the scope of this study.

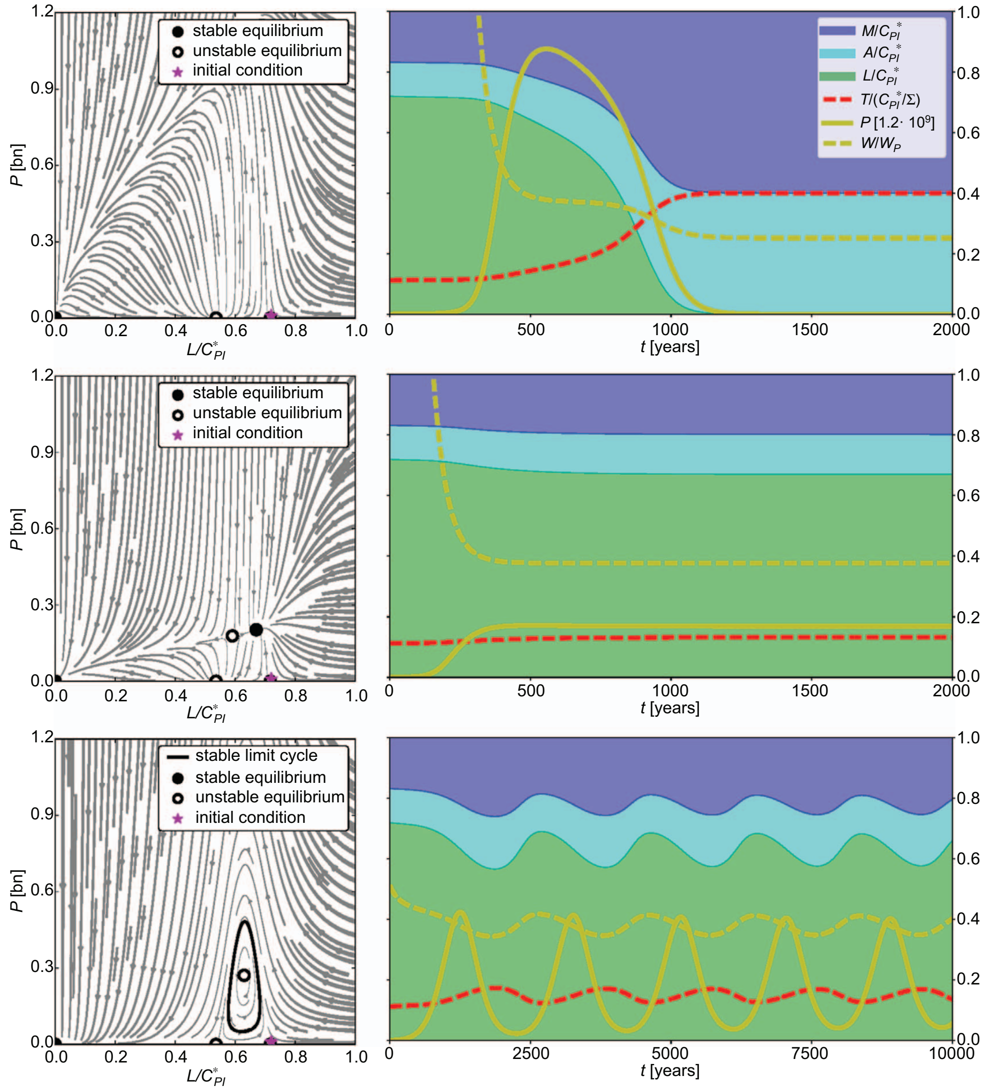


Figure 3. State space representations (left) and exemplary trajectories (right) of the non-fossil, pre-capitalistic model scenario for parameter choices giving rise to qualitatively different asymptotics of the system. In the upper panels the desert state is the only attractor, so that the population overuses natural resources and experiences a global collapse. For a lower economic productivity, shown in the middle panel, the system allows a sustainable coexistence between humans and nature, reflected by the additional attracting equilibrium in the state space. If ecosystem services are considered, an attracting limit cycle can emerge, implying sustained oscillations in all variables with a period of about 2000 years. Note the different scale of the time axis in the lower panel. All parameters but the following are set to the default values from table 1; upper panel: $y_B = 2.47{\cdot}10^{11}$ \$ GtC^{-1}, $w_L = 0$; middle panel: $y_B = 1.235{\cdot}10^{11}$ \$ GtC^{-1}, $w_L = 0$; lower panel: $y_B = 2.47{\cdot}10^{9}$ \$ GtC^{-1}, $w_L = 4.425{\cdot}10^{7}$ \$ km^2 GtC$^{-1}a^{-1}$ H^{-1}. Initial conditions: $L_0 = 2880$ GtC, $P_0 = 500000$ H.

2000 years. Qualitatively, the observed patterns are very similar to those described in classical models of predator-prey ecosystems [39]. In contrast to the latter, however, our model is still multistable in this regime since the 'desert' equilibrium is still also stable due to the functional forms for fertility and economic production. Other models of human-nature coevolution feature oscillations [15, 24, 28] which may be sustained or dampened but typically have shorter periods. The same is true for models of secular cycles [30, 40–42] which describe the emergence of oscillatory patterns due to internal socio-economic mechanisms of states or world regions.

The presented parameter settings and trajectories are of course just exemplary and hence their quantitative implications should not be overrated. There are also intermediate cases for which dampened oscillations occur, not shown here since the asymptotic states are unchanged.

The qualitative changes of the asymptotic behaviour of the system under variation of parameters can be analysed mathematically using bifurcation theory [43]. A more rigorous study reveals that there are indeed five different regimes in the (y_B, w_L) parameter space, with qualitatively different asymptotic states. However, there are only three different regimes

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

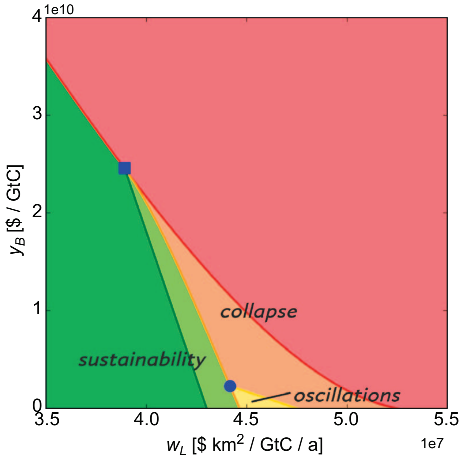


Figure 4. Bifurcation diagram in the (y_B, w_L) parameter space, showing five qualitatively different dynamic regimes. While within the two greenish regimes a *sustainable* (stable) coexistence of nature and society is possible for some initial conditions, both will *collapse* in the two reddish parameter regimes. Only the very small regime indicated in yellow features sustained *oscillations* in the dynamic variables. Borders between the regimes correspond to different local or global bifurcation curves. For details and differences within the greenish and reddish regimes, see appendix C.

(sustainability, collapse, oscillations) for which there are different *attracting* asymptotic states, as discussed above. The bifurcation diagram is shown in figure 4, the full bifurcation analysis is in appendix C.

3.3. Possible collapse of a fossil-based, capitalistic global society

We finally consider a scenario which extends the previous one in two ways. First, in addition to biomass use (B) now also fossil fuel extraction (F) from the geological pool G is enabled, where the relative shares of the two energy sources is determined by a price equilibrium. Second, physical capital K is now a stock variable with a standard growth dynamics decoupled from population growth. Altogether, this scenario applies to the era since the onset of the industrialization until recent times during which biomass and fossil fuels are the dominant energy sources and physical capital became a major factor of production. Moreover, we drop the assumption of the diffusion equilibrium from the previous scenario, giving a less stylized and more realistic representation of the global carbon cycle. Thus we have the full five-dimensional dynamical system (L, A, G, P, K) given by (2.1) to (2.5).

The availability of two different energy forms gives rise to the following question which connects closely to the introductory question of the previous section: What is the ultimate fate of the human population for different usage patterns of biomass and fossil fuels?

The proneness to use a certain form of energy is determined by various factors (see (2.8), (2.9)). It increases with the size of the associated stock variable (L for biomass, G for fossil fuels) and with the respective productivities (a_B, a_F), but decreases because of substitution effects the cheaper the other energy form is. While the stock sizes L and G are prescribed by the natural Earth system, a_B and a_F are rather abstract economical parameters which are hard to estimate from real-world data. The choice of their absolute and relative values hence facilitates an investigation of different energy usage scenarios. The oscillatory asymptotic regime discussed in section 3.2 emerged when well-being was dominated by ecosystem services. For the industrial societies considered here we assume that well-being is dominated by per-capita consumption (see the upper part in figure 4). In this part of the parameter space a variation of w_L has the same qualitative effect on the asymptotics as a variation of the economic productivity via y_B or a_B, respectively. For simplicity we subsequently choose $w_L = 0$.

To isolate the effect of emissions caused by fossil fuels, we regard a reference setting in which biomass use is disabled ($a_B = 0$) and the fossil fuel sector productivity is set to a value for which the extraction speed of fossils roughly coincides with observed values over the past 250 years ($a_F = 24.9$ GJ^{5}a^{-5} GtC^{-2} \$$^{-2}$ H^{-2}). The abundance of resources causes population and physical capital to grow fast initially until they reach a maximum after about 300 years (figure 5, upper panel). After this initial boom, well-being saturates, then both P and K slowly decrease and the economic production Y is reduced accordingly. This slow perishing of the economy and population is due to the dependence on fossil fuels from the non-renewable geological carbon stock G. After 2000 years the population is close to extinction and fossil fuels are almost depleted. Notably, for this choice of parameters, the emissions of fossil carbon only lead to a slight increase of the atmospheric carbon content (and the associated global mean temperature), while most of the carbon is captured in biomass and soils. Also for other values of a_F, a collapse of the terrestrial system to a desert state due to emissions of fossil fuels is not observable in the model. However, the fate of a population in this purely fossil-based scenario is slow extinction on a well-forested planet, but now with an almost unchanged level of well-being until the end.

Obviously, this scenario is not very realistic since humans would certainly start to (and historically always did) harvest biomass in order to satisfy their need for energy. By choosing a rather low biomass sector productivity of $a_B = 0.05\ a_F$ the initial share of biomass in total energy use amount to about 15% (figure 5, middle panel). The behaviour of the system during the first 500 years of simulation time is very similar to the reference setting with the only difference that, due to the additional use of biomass, P, K and thus Y reach higher absolute levels. Due to the depletion of the geological carbon stock and the increase in terrestrial carbon, the share of biomass is constantly increasing and overtakes the fossil share

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

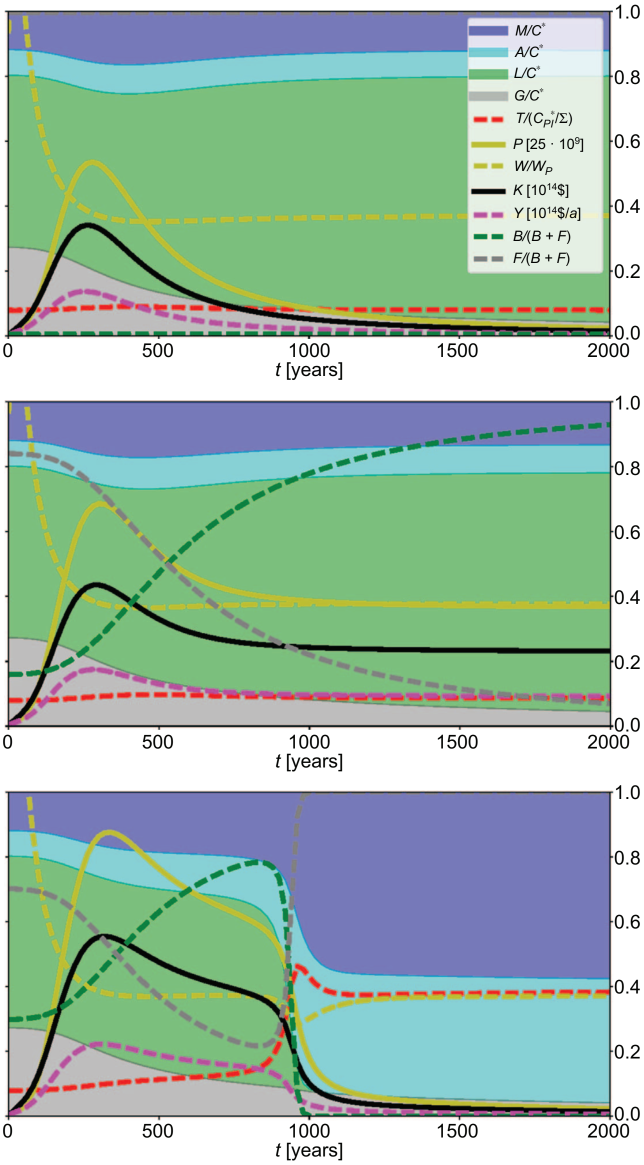


Figure 5. Exemplary trajectories of the fossil-based, capitalistic model scenario for different usage of biomass and fossil fuels reflected by different combinations of the sector productivities a_B and a_F. In the fossil-only reference setting (upper panel) the global will go extinct after several millennia with the depletion of the geological carbon stock while the emitted carbon is mainly stored in the terrestrial stock. Moderate usage of biomass allows a sustained coexistence of humans and nature in the long run (middle panel) but fossil resources will still be completely depleted. When humans exert too much pressure on the terrestrial system through biomass use (land use) these can ultimately collapse, thereby ruining the preconditions for life on Earth (lower panel). The socio-economic development is indistinguishable in the scenarios with enabled biomass use until about 800 years of simulation time. Only changing the continued changes in the natural subsystem of Earth indicate the prolonged transient towards an undesirable desert state. All parameters but the following are set to the default values from table 1; upper panel: $a_F = 24.9\ \mathrm{GJ^5 a^{-5}\ GtC^{-2}\ \$^{-2}\ H^{-2}}$, $a_B = 0$; middle panel: $a_F = 24.9\ \mathrm{GJ^5 a^{-5}\ GtC^{-2}\ \$^{-2}\ H^{-2}}$, $a_B = 1.25\ \mathrm{GJ^5 a^{-5}\ GtC^{-2}\ \$^{-2}\ H^{-2}}$; lower panel: $a_F = 24.9\ \mathrm{GJ^5 a^{-5}\ GtC^{-2}\ \$^{-2}\ H^{-2}}$, $a_B = 2.8\ \mathrm{GJ^5 a^{-5}\ GtC^{-2}\ \$^{-2}\ H^{-2}}$. Initial conditions: $L_0 = 2915$ GtC, $P_0 = 162{\cdot}10^6$ H, $K_0 = 323{\cdot}10^9$ \$.

after about 500 years. In contrast to the previous setting the global society has an alternative to fossil fuels and is not doomed to go extinct. Instead, the population decrease slows down and a sustained coexistence between humans and nature emerges. Note that humans still continue to use fossil fuels until ultimately the geological carbon stock is completely depleted, which follows from the economical model of the energy sector ($F = 0$ necessitates $G = 0$ as long as $P, K > 0$, see (2.9)). An abandoning of fossil fuel use can thus not be achieved by the economic forces assumed in the model; instead this would necessitate other economical mechanisms, e.g. banning or taxing of fossils through policies. In the asymptotic state about 10 bn humans inhabit Earth, the average per-capita capital amounts to about 2500 \$ which we regard as realistic orders of magnitude.

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

So can we conclude that biomass use can save humankind when fossils are abandoned for whatever reason? This must clearly be denied as our last parameter setting shows, in which assume a larger biomass sector productivity ($a_B = 0.1125\ a_F$). Now biomass initially makes up about a third of the total energy used and becomes the dominant form of energy after about 350 years. Again the socio-economic observables (P, K, Y) behave qualitatively very similar to the previous settings (fast increase to a maximum, followed by slowing decrease) until about 800 years of simulation time. About this time their speed of decrease accelerates again and they drop to very low values within about 200 years. This breakdown of the socio-economic system is caused by overuse of natural resources which triggered a collapse of the biosphere (represented by the terrestrial carbon stock L) to the desert state, just as observed in the non-fossil, pre-capitalistic scenario discussed in section 3.2. After the collapse humans can only 'survive' until the remaining fossil fuel resources are completely depleted, so that ultimately, an unpopulated desert planet prevails. This is, of course, not realistic for several reasons: the life-enabling capacity of the biosphere (e.g. through oxygen production) is not accounted for and renewable energy is not available in the model. We thus learn that the intensity of biomass and land use, reflected by the parameter a_B are of crucial importance for a sustainable global coevolution of humans and nature which should always be considered besides the necessity for reducing emissions from fossil fuels. While the parameter value is fixed in the model simulation, in reality, the socio-economic conditions it reflects can be subject to change, e.g. through policy instruments.

It should be pointed out that the collapse of the system in the third setting could not have been predicted by looking solely at socio-economic observables, as these evolve analogously in the previous settings for roughly the first 800 years of simulation time. Merely the changing environmental conditions, as indicated by the continued increase in global mean temperature and decrease of the vegetation from year 300 to 800, qualitatively differentiate this setting from the previous ones and thus hint at the fact that the system actually undergoes a long transient period towards an undesirable final state. Note that we do not even need to model direct climate damages on, say, mortality and capital depreciation, to cause the extinction.

A second question posed by the industrialization scenario is: What is the effect of the dynamic physical capital stock K, compared to the non-capitalistic societies discussed above? For all regarded parameter settings population and capital evolve alike, meaning a constant capital per capita just as it was assumed in the previous, non-capitalistic scenario. This observation can be explained with the rate of capital depreciation (k) which is comparable to the reproduction rates of humans. A considerably lower depreciation rate would instead introduce a time lag between the trajectories of P and K. The estimated parameters, however, indicate rather short time scales for the changes of the factors of production, compared to the rather slow evolution of the carbon stocks (apart from collapses).

4. Conclusions

We presented a flexible conceptual World-Earth model which is—through an appropriate choice of variables and parameters—able to qualitatively represent the global coevolutionary dynamics of humans and nature for different socio-cultural stages of human history on Earth, particularly during the Holocence and Anthropocene epochs. The actual evolution of global carbon stocks was found to oppose the dynamics to be expected from the topology of the natural carbon cycle, which is mainly due to human interference with natural dynamics through land use (change) and emissions of carbon into the atmosphere. Due to various nonlinearities in natural and social dynamics, an accurate description of the mid and long-term evolution of the Earth system thus necessitates an explicit modelling of the 'human factor' with a balanced representation of natural and socio-economic subsystems. Our conceptual model (framework) thus contributes to the challenge of 'Modelling the Anthropocene'.

For each model scenario we identified the characteristics of possible asymptotic states of the system which comprise a sustainable coexistence of humans and nature, a collapse of both natural and socio-economic subsystems and even persistent oscillatory dynamics with multi-millennial periods. By systematic variation of those parameters whose estimates from real-world data are particularly uncertain, we found the preconditions of the different asymptotic patterns. It is especially those parameters related to the appraisal (w_L) or the intensity of use (y_B, a_B) of the biosphere, which make a crucial difference for the fate of the planet and humankind.

The overall picture of our results supports the insight that neither fossil fuels nor biomass use are likely to facilitate a sustainable coexistence of several billion humans on a planet with limited natural resources. We conclude that besides reducing the global demand for energy, merely the extensive use of renewable energy forms may pave the way into a sustainable future of a well-developed global society. Extending the current framework by enabling the use of renewables is thus a priority for the future model development.

In our model analysis we focussed mainly on understanding the asymptotic behaviour of the coevolutionary Earth system and hence regarded rather long time scales of several centuries to

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

millennia. A lot of interesting dynamics like growth phases or collapses, can, however, happen on quite short time scales from decades to centuries. These transient phases could reveal interesting insights, particularly regarding the evolution of the socio-economic subsystem of the Earth. We believe that historically observed phenomena like the 'Great Acceleration' [44] could, in principle, be reproduced with our model, given appropriate parameter values and initial conditions. To show this, in appendix D we derive conditions under which the socio-economic observables of the model (K, Y, P) feature super-exponentially fast growth. An interesting extension would be to replace the global society of our model by a number of interacting regional societies. One could then also add socio-cultural model components describing warfare, internal conflicts, or the level of social and political order ([30]) and thus study the interaction between slower global cycles and faster domestic cycles.

Beyond the implications for global sustainability our simple model studies emphasize the subtleties resulting from the nonlinear characteristics of the Earth system, e.g. depicted by very long-lasting transients towards undesirable attractors. Realizing that such dynamical features can even emerge in simple conceptual models like the presented ones, should raise the awareness and caution also for the analysis of more comprehensive models of the Earth system.

Acknowledgments

We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the Göttingen University.

We further thank Wolfram Barfuss, Jonathan Donges, Vera Heck, Tim Kittel and the whole COPAN group at PIK for helpful discussions and comments throughout all stages of this work.

Appendix A: Derivation of the model

Variables

The two main variables of interest for this model are human well-being W (representing the most important aspect of the anthroposphere or socio-economic subsystem of Earth) and terrestrial carbon stock L (representing the most important aspect of the ecosphere or biophysical subsystem of Earth). We try to restrict the model to those further variables and processes that seem indispensable in order to assess the qualitative features of the possible coevolutionary pathways of L and W on a time-scale of hundreds to thousands of years, hence we include the following quantities needed to represent a carbon cycle and resource-dependent economic and population growth:

- Time t [standard unit: years, a].
- Terrestrial ('land') carbon stock $L \in [0, C^*]$ [GtC] (including soil and plants).
- Atmospheric carbon stock $A \in [0, C^*]$ [gigatons carbon, GtC].
- Accessible geological carbon stock (serving as fossil fuel reserves) $G \in [0, C^*]$ [GtC].
- Maritime carbon stock $M = C^* - A - G - L \in [0, C^*]$ [GtC] (including only the upper part of the oceans which exchanges carbon comparatively fast with air).
- Human population stock $P \geq 0$ [number of humans, H].
- Physical capital stock $K \geq 0$ [time-independent (e.g. 2011) US dollars, $].
- Global mean surface air temperature $T \geq 0$ (representing 'climate'), measured not in Kelvin but for simplicity in 'carbon-equivalent degrees' [Ced = GtC], using an atmospheric carbon-equivalent scale. i.e. $T = x$ Ced is the equilibrium temperature of an atmosphere containing x GtC).
- Biomass extraction flow $B \geq 0$ [GtC/a] and biomass energy flow $E_B \geq 0$ [GJ/a].
- Fossil carbon extraction flow $F \geq 0$ [GtC/a] and fossil energy flow $E_F \geq 0$ [GJ/a].
- Total energy input flow $E \geq 0$ [GJ/a].
- Economic output flow $Y \geq 0$ [$/a].
- Investment flow $I \geq 0$ [$/a].
- Well-being W in per-capita consumption-equivalent units [$/a H] (including economic welfare and environmental effects, e.g. health and ecosystem services).

We follow the predominant economic convention of measuring capital, production, and consumption in monetary units. A, B, E, F, G, I, K, L, M, P, Y are *extensive* quantities in the sense that the would double if the Earth System was replaced by two identical copies of itself, while T and W are *intensive* quantities which would not double. The only *conserved* quantity in the model is carbon, as expressed by the equation $A + G + L + M \equiv C^*$.

Processes, generic interaction terms and equations

The following processes and dependencies are considered to be the main drivers of the carbon cycle, economic and population growth:

- Ocean to air diffusion $f_{\text{diff.}}(A, M)$ [GtC/a] (ignoring pressure and temperature dependency).

- Greenhouse effect on temperature[7] $T = T(A)$ [GtC] (ignoring other GHG).
- Land to air respiration $f_{\text{resp.}}(L, T) \geq 0$ [GtC/a] (ignoring other dependencies).
- Photosynthesis $f_{\text{photos.}}(A, L, T) \geq 0$ [GtC/a] (ignoring nitrogen and other dependencies).
- Biomass extraction $B = B(G, K, L, P) \geq 0$ and combustion $E_B = E_B(B)$ (ignoring other economic dependencies, and afforestation, carbon storage and other policy dependencies, and assuming almost all extracted land carbon ends up in the atmosphere after a negligible time; ignoring carbon stored in human bodies and physical capital).
- Fossil fuel extraction $F = F(G, K, L, P) \geq 0$ and combustion $E_F = E_F(G)$.
- Total energy usage from these energy sources $E = E_B + E_F$.
- Economic production of output $Y = Y(E, K, P)$ (assuming the two energy sources are perfect substitutes).
- Capital growth through investment $I = iY$.
- Capital depreciation $f_{\text{deprec.}}(K) \geq 0$ [\$/a].
- Consumption of all non-invested economic output and emergence of well-being $W = W(L, P, Y)$.
- Population fertility and mortality $f_{\text{fert./mort.}}(W)$ [1/a].

This leads to the following generic equations:

$$dL/dt = f_{\text{photos.}}(A, L, T) - f_{\text{resp.}}(L, T) - B, \tag{A.1}$$

$$dA/dt = -dL/dt + F + f_{\text{diff.}}(A, M), \tag{A.2}$$

$$dG/dt = -F, \tag{A.3}$$

$$dK/dt = iY - f_{\text{deprec.}}(K), \tag{A.4}$$

$$dP/dt = (f_{\text{fert.}} - f_{\text{mort.}})(W)P \tag{A.5}$$

with

$$T = T(A), \tag{A.6}$$

$$B = B(G, K, L, P), \tag{A.7}$$

$$F = F(G, K, L, P), \tag{A.8}$$

$$E = E_B(B) + E_F(F) \tag{A.9}$$

$$Y = Y(E, K, P), \tag{A.10}$$

$$W = W(L, P, Y). \tag{A.11}$$

Choice of functional forms

Since our aim is a mainly qualitative analysis rather than quantitative prediction, we aim at choosing simple functional forms that fulfil at least the following qualitative properties:

- $f_{\text{diff.}}$ is increasing in M and decreasing in A.
- T is increasing in A.
- $f_{\text{resp.}}$ is roughly proportional to L and is increasing but concave in T (over the range of temperatures experienced in the holocene).
- $f_{\text{photos.}}$ is roughly proportional to L, is increasing and concave in A (due to diminishing marginal crabon fertilization), and is decreasing in T (over the range of temperatures experienced in the holocene).
- $f_{\text{deprec.}}$ is roughly proportional in K.
- $f_{\text{fert.}}$ is zero for vanishing W, grows roughly proportionally with W for small values of W (representing basic nutritional needs for reproduction as in ecological models), grows more concavely when W grows further until W reaches some value $W_P > 0$ (representing saturation of fertility due to biological limits) and finally declines again towards zero when W grows even further (due to education- and social security-related effects).
- $f_{\text{mort.}}$ is infinite for vanishing W and declines towards zero with growing W.
- $E_B, E_F \geq 0$ are roughly proportional to B or F, respectively.
- B is increasing in K, L due to lower costs, increasing in P due to higher demand, and convexly decreasing in G due to substitution by fossil fuel. Analogously, F is increasing in G, K, P and convexly decreasing in L.
- Y is increasing and concave in all of E, K, P.

We fulfil most of these by the following simple choices:

- $f_{\text{diff.}}(A, M) = d(M - mA)$.
- $T = A/\Sigma$ (T is measured in carbon-equivalent degrees and an *intensive* quantity).
- $f_{\text{photos.}}(A, L, T) = (l_0 - l_T T)\sqrt{A/\Sigma}L$[8].
- $f_{\text{resp.}}(L, T) = (a_0 + a_T T)L$.
- $f_{\text{deprec.}}(K) = kK$.

[7] A model version in which T is a state variable with a transient response to atmospheric carbon A has been studied. As it reveals the same asymptotic behaviour and the estimated timescale of the response is rather fast, we assume for this study the greenhouse effect to be instantaneous.

[8] The exponent 1/2 for A in the fertilization term is larger but simpler than the choice of 0.3 in [31].

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

- $W(L, P, Y) = (1 - i)Y/P + w_L L$ with ecosystem services coefficient w_L.
- $f_{\text{fert.}}(W) = 2pW_P W/(W_P^2 + W^2)$ with a maximum fertility of $p > 0$ reached at the saturation well-being level $W_P > 0$.
- $f_{\text{mort.}}(W) = q/W$ with mortality coefficient $q > 0$.
- $E_B = e_B B$ and $E_F = e_F F$ with combustion efficiencies $e_B, e_F > 0$.

The formulae for B, F, Y are derived from the following economic submodel.

Two-sector economic submodel

We assume the global economy produces output using a global production function

$$Y = f(P, K, L, G),$$

using P as a source of labour and L, G as sources of energy. In the full model, we assume larger population numbers lead to increasing globalization with overall positive effects on productivity, hence we will aim at choosing an f that has increasing returns to scale, i.e. $f(aP, aK, aL, aG) > af(P, K, L, G)$ for all $a > 1$. In the reduced model for pre-capitalistic societies, we will keep the more traditional assumption of constant returns to scale, i.e. $f(aP, aK, aL, aG) = af(P, K, L, G)$ for all $a > 1$. This will influence our choice of elasticities (see below). In order to be able to model substitution effects between the two different resource use flows B and F, we need to distinguish the energy sector(s) from the rest of the economy (which we call the 'final' sector). A quite general modelling approach for doing this is to assume nested production functions

$$Y = f(P, K, L, G) = f_Y(P_Y, K_Y, E_B, E_F),$$

$$E_B = f_B(P_B, K_B, L),$$

$$E_F = f_F(P_F, K_F, G)$$

and determine the unknown labour and capital shares $P_\cdot, K_\cdot$ by some form of social optimization or market mechanism. Since this will in general lead to quite complicated expressions for Y, E_B, E_F, we make a number of strong simplifying and symmetry assumptions here in order to get manageably simple formulae.

To reduce the number of independent factors in f, we treat the two energy forms as perfect substitutes, so that $Y = f_Y(P_Y, K_Y, E)$ with total energy input $E = E_B + E_F$. Since energy is generally considered an input that cannot be substituted well by other factors, the natural candidate to model the dependency of Y on E is not a CES production function but either a Cobb-Douglas or a Leontieff production function. We choose the simpler, a Leontieff form, which amounts to prescribing a fixed ratio of energy need per output that is independent of the other factors:

$$Y = y_E \ \min\{E, g_Y(K_Y, P_Y)\},$$

where $y_E > 0$ is an energy productivity factor (the inverse of the final sector's energy intensity). We assume the standard Cobb-Douglas form for the relative substitutability of labour and capital:

$$g_Y(K_Y, P_Y) = b_Y K_Y^{\kappa_Y} P_Y^{\pi_Y}$$

with productivity $b_Y > 0$ and elasticities $0 < \kappa_Y, \pi_Y < 1$. In each of the two forms of energy, we also assume the Cobb-Douglas form,

$$E_B = b_B K_B^{\kappa_B} P_B^{\pi_B} L^\lambda,$$

$$E_F = b_F K_F^{\kappa_F} P_F^{\pi_F} G^\gamma,$$

with sectoral productivities $b_B, b_F > 0$ and further elasticities $\kappa_\cdot, \pi_\cdot, \lambda, \gamma$.

Although the simplest assumption about the allocation of labour and capital to the three production processes f_Y, f_B, f_F would be to assume fixed shares, this would ignore the strong incentive to allocate the resources to the production of the more productive energy form, and to allocate the more resources to energy production the more productive the energy sector is compared to the final sector. The next-best simple assumption is a social planner perspective that allocates resources so as to maximize final output Y. We prefer this to the alternative view of a competitive allocation via factor markets for two reasons: (i) the latter view is more closely tied to the assumption of a specific economic system, which is less plausible for the long time horizons we aim at, and (ii) if markets are approximately perfect, they would lead to maximizing final output anyway.

To get this solution, we first assume the energy sector's inputs K_E, P_E were known and solve the intra-energy-sector allocation problem via the first-order conditions

$$\partial E_B/\partial K_B = \partial E_F/\partial K_F, \partial E_B/\partial P_B = \partial E_F/\partial P_F$$

under the constraints

$$K_B + K_F + K_R = K_E, P_B + P_F + P_R = P_E.$$

It turns out that this only leads to sufficiently simple expressions if we assume that the labour elasticities π_B, π_F of the two energy forms are equal, and similarly for capital, hence we put $\kappa_{B,F} \equiv \kappa_E$ and $\pi_{B,F} \equiv \pi_E$ and get

$$K_B = X_B K_E/X_E, K_F = X_F K_E/X_E,$$

$$P_B = X_B P_E/X_E, P_F = X_F P_E/X_E,$$

$$E_B = X_B Z_E, E_F = X_F Z_E,$$

where

$$X_B = b_B^{\alpha_E} L^{\alpha_E \lambda}, X_F = b_F^{\alpha_E} G^{\alpha_E \gamma},$$

$$X_E = X_B + X_F,$$

$$Z_E = K_E^{\kappa_E} P_E^{\pi_E} / X_E^{\kappa_E + \pi_E},$$

$$\alpha_E = 1/(1 - \kappa_E - \pi_E).$$

Given K_E, P_E, we thus have

$$E = X_E Z_E = K_E^{\kappa_E} P_E^{\pi_E} X_E^{1/\alpha_E}.$$

Since neither the energy nor the final sector are to have idle resources, we must also have

$$E = g_Y(K_Y, P_Y) = b_Y K_Y^{\kappa_Y} P_Y^{\pi_Y}.$$

An optimal allocation between energy and final sector then requires that no 'trade' in capital or labour is profitable beween the two sectors, which in view of the constraint $E = g_Y$ leads to the additional equation

$$\frac{\partial g_Y / \partial K_Y}{\partial E / \partial K_E} = \frac{\partial g_Y / \partial P_Y}{\partial E / \partial P_E},$$

i.e.

$$\frac{\kappa_Y g_Y / K_Y}{\kappa_E E / K_E} = \frac{\pi_Y g_Y / P_Y}{\pi_E E / P_E}$$

which implies

$$\frac{\kappa_Y K_E}{\kappa_E K_Y} = \frac{\pi_Y P_E}{\pi_E P_Y} =: \beta.$$

To find β, we solve

$$\begin{aligned} 0 &= E - g_Y \\ &= \left(\frac{\beta \kappa_E K_Y}{\kappa_Y}\right)^{\kappa_E} \left(\frac{\beta \pi_E P_Y}{\pi_Y}\right)^{\pi_E} X_E^{1/\alpha_E} - b_Y K_Y^{\kappa_Y} P_Y^{\pi_Y} \end{aligned}$$

and get

$$\beta^{\kappa_E + \pi_E} = b_Y \left(\frac{\kappa_Y}{\kappa_E}\right)^{\kappa_E} \left(\frac{\pi_Y}{\pi_E}\right)^{\pi_E} K_Y^{\kappa_Y - \kappa_E} P_Y^{\pi_Y - \pi_E} X_E^{-1/\alpha_E}.$$

We note that this simplifies considerably if for each of the factors capital and labour, either only one of the sectors requires it or both sectors have the same elasticity for it. Since clearly a considerable amount of capital and labour are needed in both sectors, we hence assume $\kappa_E = \kappa_Y =: \kappa$ and $\pi_E = \pi_Y =: \pi$. We can now solve

$$\frac{K_E}{K - K_E} = \frac{P_E}{P - P_E} = \beta = (b_Y X_E^{-1/\alpha_E})^{1/(\kappa+\pi)},$$

$$K_E = \frac{\beta}{1+\beta} K, P_E = \frac{\beta}{1+\beta} P,$$

$$K_Y = \frac{1}{1+\beta} K, P_Y = \frac{1}{1+\beta} P.$$

Putting all of the above together, using $\eta = 1/(1 + 1/\beta)$ (the share of the energy sector) instead of β, and introducing $\alpha = 1/(1 - \kappa - \pi)$, $a_B = b_B^\alpha$ and $a_F = b_F^\alpha$, we get

$$X_B = a_B L^{\alpha\lambda}, K_B = \frac{X_B}{X} K_E, P_B = \frac{X_B}{X} P_E,$$

$$X_F = a_F G^{\alpha\gamma}, K_F = \frac{X_F}{X} K_E, P_F = \frac{X_F}{X} P_E,$$

$$X = X_B + X_F, \eta = \frac{1}{1 + (X^{1/\alpha}/b_Y)^{1/(\kappa+\pi)}},$$

$$Z = K_E^\kappa P_E^\pi / X^{\kappa+\pi} = \eta^{\kappa+\pi} K^\kappa P^\pi / X^{\kappa+\pi},$$

$$E = XZ, K_E = \eta K, P_E = \eta P,$$

$$Y = y_E E, K_Y = (1-\eta) K, P_Y = (1-\eta) P,$$

$$Z' = \left(1 + \frac{(a_B L^{\alpha\lambda} + a_F G^{\alpha\gamma})^{\frac{1-\kappa-\pi}{\kappa+\pi}}}{b_Y^{1/(\kappa+\pi)}}\right)^{-\kappa-\pi},$$

$$E_B = X_B Z = \frac{a_B L^{\alpha\lambda} K^\kappa P^\pi}{(a_B L^{\alpha\lambda} + a_F G^{\alpha\gamma})^{\kappa+\pi}} Z',$$

$$E_F = X_F Z = \frac{a_F G^{\alpha\gamma} K^\kappa P^\pi}{(a_B L^{\alpha\lambda} + a_F G^{\alpha\gamma})^{\kappa+\pi}} Z'.$$

For the economy to have increasing returns to scale, we choose elasticities that fulfil $\kappa + \pi + \min(\lambda, \gamma) > 1$. A simple choice which is roughly in line with estimates of labour and capital elasticities in the agricultural sector of many countries is $\kappa = \pi = \lambda = \gamma = 2/5$. Then $\kappa + \pi = 4/5$, $\alpha = 5$, $\alpha\lambda = \alpha\gamma = 2$, and hence

$$E_B = \frac{a_B L^2 (PK)^{2/5}}{(a_B L^2 + a_F G^2)^{4/5}} \left(1 + \frac{(a_B L^2 + a_F G^2)^{1/4}}{b_Y^{5/4}}\right)^{-4/5},$$

$$E_F = \frac{a_F G^2 (PK)^{2/5}}{(a_B L^2 + a_F G^2)^{4/5}} \left(1 + \frac{(a_B L^2 + a_F G^2)^{1/4}}{b_Y^{5/4}}\right)^{-4/5}.$$

Finally, we assume that $b_Y^5 \gg a_B L^2 + a_F G^2$ so that the share of the energy sector η (the large bracket) is ≈ 1. Note that as the 'energy' sector in our model includes all of agriculture, a very large share of this sector is not too implausible. We thus arrive at the simple approximation used in the model,

$$B = \frac{a_B}{e_B} \frac{L^2 (PK)^{2/5}}{(a_B L^2 + a_F G^2)^{4/5}},$$

$$F = \frac{a_F}{e_F} \frac{G^2 (PK)^{2/5}}{(a_B L^2 + a_F G^2)^{4/5}},$$

$$Y = y_E (e_B B + e_F F).$$

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

For the pre-capitalistic variant of the model, we choose $\kappa = \lambda = 3/10$ instead to get constant returns to scale. Together with a fixed per capita capital of $K \propto P$, this gives equations (3.1) and (3.2).

Appendix B: Parameter estimation

The available Earth surface area (Σ) has been identified with the Earth's current land surface area. The parametrization of the carbon cycle parameters (C^*, C^*_{PI}, a_0, a_T, l_0, l_T, d, m) occurred on the basis of the recent estimated of carbon stocks and flows by the International Panel on Climate Change [36]. The estimates of the demographic parameters (p, W_P, q) result from separately performed weighted least squares regressions of the modelled dependencies of fertility and mortality on well-being (equation (2.4)), respectively. As input data we used estimates of various World Development Indicators for which country-wise, yearly data are available from the World Bank [45]. The investment rate (i) has been estimated by averaging the global times series on 'gross capital formation' by the World Bank [45]. A reasonable value for the capital depreciation rate (k) can be found in [46]. Typical energy densities of biomass (e_B) and fossil fuels (e_F) are of comparable size [47]. The economic output per (primary) energy input has been estimated as the average of the inverse of the time series on 'energy intensity level of primary energy' available from the World Bank [45].

The subsequently introduced parameters y_B and b in the non-fossil scenario (section 3.2) have been estimated using data on global population level, agricultural sector's value added to the gross world product and the contribution of harvesting to the 'Human Appropriation of Net Primary Production' (HANPP) [45, 48].

Appendix C: Bifurcation analysis

The rather low-dimensional complexity and the simple functional relationships (see equations (2.1) to (2.5)) of the presented model facilitate the application of analysis techniques from dynamical systems theory, e.g. bifurcation analysis [43]. Bifurcation analysis aims at a partition of a dynamical system's parameter space into regimes, such that within different regimes the system's state spaces are topologically non-equivalent, meaning different numbers or stabilities of the system's equilibria or limit cycles and hence a different asymptotic behaviour.

For this work we conducted a bifurcation analysis of the (y_B, w_L)-parameter-subspace of the two-dimensional (L, P) submodel discussed in section 3.2. The bifurcation diagram in figure 4) shows a partition of the parameter space into five regimes for which the corresponding state spaces are topologically non-equivalent. The borders between the regimes correspond to codimension-1-bifurcations, while the blue points at their intersections indicate bifurcations of codimension 2.

Suppose the parameter values lie within the large reddish region in figure 4 for which the 'desert' state is the only attractor of the system. When crossing the red curve above the blue square, the system undergoes a (local) *fold* (or *saddle-node*) bifurcation leading to the existence of an unstable (saddle) equilibrium and a stable (node) equilibrium in the dark green regime which hence facilitates a sustainable coexistence of humans with nature. Crossing the green curve gives rise to a (global) homoclinic bifurcation through which an unstable limit-cycle is created. However, this does not alter the set of attractors, hence the qualitative asymptotics remain unchanged. If the orange curve is transgressed from within the light green region, an *Andronov-Hopf* bifurcation occurs. It is *sub-critical* when the curve is crossed above the blue circle. In this case the unstable limit-cycle coalesces with the stable node, leaving an unstable node in the orange region. When the orange curve is crossed below the blue circle, the Andronov-Hopf bifurcation is *super-critical*, meaning that a stable limit-cycle is born around the stable coexistence equilibrium which in turn becomes unstable. The yellow region hence features an attracting limit-cycle besides the stable desert state. The yellow bifurcation curve corresponds to a *fold bifurcation of cycles* in which the two limit-cycles coalesce and vanish, leaving an unstable node in the orange region. Hence, in the orange regime the systems features a saddle point and an unstable node with $P > 0$, which undergo a fold bifurcation when the red line is crossed from left to right below the blue square. In the orange and red regions the desert state is the only attractor, meaning that ultimately nature and society are doomed to collapse.

At the point marked by the blue square at which the fold, Andronov-Hopf and homoclinic bifurcation curves intersect, a so-called *Bogdanov-Takens* bifurcation occurs. The point marked by the blue square at which the fold-of-cycles curve connects to the two branches of the Andronov-Hopf curve is referred to as a *Bautin* (or *generalized Hopf*) bifurcation.

Note that in figure 4 only the fold and Andronov-Hopf curves which correspond to *local* bifurcations have been computed numerically, using the software PyDSTool [49]. As the tool is not able to detect *global* bifurcations, the homoclinic and fold-of-cycles curves, whose existence is known from theory, are indicated only schematically.

Appendix D: Conditions for superexponential growth

Due to several nonlinearities in our model, most quantities can show both sub- and superexponential growth or decay, in contrast to most basic purely economic growth models.

A quantity x has a phase of superexponential growth whenever $0 < d^2(\ln x)/dt^2 = (\ddot{x}x - \dot{x}^2)/x^2$.

For population P, we have $d(\ln P)/dt = \dot{P}/P = f(W) := \frac{2WW_P}{W^2+W_P^2}p - \frac{q}{W}$ and f is negative if $0 < W < W_0$ (for some constant W_0), positive and increasing if $W_0 < W < W^*$, and positive and decreasing if $W^* < W$, where $0 < W_0 < W_P < W^*$. Hence P has superexponential growth iff either (i) $W_0 < W < W^*$ and $\dot{W} > 0$, or (ii) $W^* < W$ and $\dot{W} < 0$, i.e. when well-being is moving towards the point where net reproduction is maximal.

For capital K, the condition is

$$\begin{aligned}
0 &< \ddot{K}K - \dot{K}^2 \\
&= K\frac{d}{dt}(iy_B(a_BL^2 + a_FG^2)^{1/5}(PK)^{2/5} - kK) - \dot{K}^2 \\
&= K(iy_E(a_BL^2 + a_FG^2)^{1/5}(PK)^{2/5} \\
&\quad \times \left(\frac{2a_BL\dot{L} + 2a_FG\dot{G}}{5(a_BL^2 + a_FG^2)} + \frac{2\dot{P}}{5P} + \frac{2\dot{K}}{5K}\right) - k\dot{K}) - \dot{K}^2 \\
&= K\left((\dot{K} + kK)\frac{2}{5}\left(\frac{a_BL\dot{L} + a_FG\dot{G}}{a_BL^2 + a_FG^2} + \frac{\dot{P}}{P} + \frac{\dot{K}}{K}\right)\right. \\
&\quad \left. - k\dot{K}\right) - \dot{K}^2.
\end{aligned}$$

If $\dot{K} > 0$, this condition is the more likely fulfilled the smaller $\dot{K}$, L, G, and P, and the larger K, $\dot{L}$, $\dot{G}$, $\dot{P}$, and k. Hence a small l_T, a_0, a_T, i, y_E, a_B, a_F, q, and q_P, a large A, l_0, e_B, e_F, and p, and a $W \approx W_P$ tend to make a superexponential growth of K more likely.

References

[1] Crutzen P J 2002 *Nature* **415** 23

[2] Zalasiewicz J, Williams M, Haywood A and Ellis M 2011 *Phil. Trans. R. Soc.* A **369** 835–41

[3] Malm A and Hornborg A 2014 *The Anthropocene Review* **1** 62–9

[4] Lewis S L and Maslin M A 2015 *Nature* **519** 171–80

[5] Waters C N *et al* 2016 *Science* **351** aad2622-1–10

[6] Schellnhuber H J 1998 Discourse: Earth System Analysis—The Scope of the Challenge, *Earth System Analysis: Integrating Science for Sustainability* ed H J Schellnhuber and V Wenzel (Berlin: Springer) 3–195

[7] Schellnhuber H J 1999 *Nature* **402** C19–23

[8] Verburg P H, Dearing J A, Dyke J G, van der Leeuw S, Seitzinger S, Steffen W and Syvitski J 2016 *Glob. Environ. Change* **39** 328–40

[9] van Vuuren D P, Lucas P L, Häyhä T, Cornell S E and Stafford-Smith M2016 *Earth Sys. Dyn.* **7** 267–79

[10] Moss R *et al* 2008 *Towards New Scenarios for Analysis of Emissions, Climate Change, Impacts, and Response Strategies* Technical summary (Geneva: Intergovernmental Panel on Climate Change) p 25

[11] Weyant J, Davidson O, Dowlabathi H, Edmonds J, Grubb M, Parson E A, Richels R, Rotmans J, Shukla P R and Tol R S J 1996 Integrated assessment of climate change: an overview and comparison of approaches and results *Climate Change 1995: Economic and Social Dimensions of Climate Change. Contribution of Working Group III to the Second Assessment Report of the Intergovernmental Panel on Climate Change* ed J P Bruce, E F Haites and H Lee (Cambridge: Cambridge University Press) ch 10 pp 367–96

[12] Meadows D H, Meadows D L, Randers J and Behrens W W III 1972 *The Limits to Growth* (New York: Universe Books)

[13] Meadows D H, Randers J and Meadows D L 2004 *Limits to Growth: The 30-Year Update* (White River Junction, VT: Chelsea Green Publishing)

[14] Boumans R, Costanza R, Farley J, Wilson M A, Portela R, Rotmans J, Villa F and Grasso M 2002 *Ecol. Eco.* **41** 529–60

[15] Motesharrei S, Rivas J and Kalnay E 2014 *Ecol. Eco.* **101** 90–102

[16] Kittel T, Koch R, Heitzig J, Deffuant G, Mathias J D and Kurths J 2017 Operationalization of topology of sustainable management to estimate qualitatively different regions in state space in preparation

[17] Rockström J *et al* 2009 *Ecol. Soc.* **14** 32

[18] Rockström J *et al* 2009 *Nature* **461** 472–75

[19] Steffen W *et al* 2015 *Science* **347** 1259855

[20] Raworth K 2012 *Oxfam Policy and Practice: Climate Change and Resilience* **8** 1–26

[21] United Nations General Assembly 2015 *Transforming our world: the 2030 agenda for sustainable development* (UN General Assembly) A/RES/70/1

[22] Folke C, Biggs R, Norström A V, Reyers B and Rockström J 2016 *Ecol. Soc.* **21** 41

[23] Heitzig J, Kittel T, Donges J F and Molkenthin N 2016 *Earth Sys. Dyn.* **7** 21–50

[24] Brander J A and Taylor M S 1998 *Am. Eco. Rev.* **88** 119–38

[25] Brandt G and Merico A 2013 *Ecol. Com.* **13** 46–52

[26] Wiedermann M, Donges J F, Heitzig J, LuchtWand K J 2015 *Phy. Rev.* E **91** 052801

[27] Barfuss W, Donges J F, Wiedermann M and Lucht W 2017 *Earth Syst. Dyn.* **8** 255–64

[28] Kellie-Smith O and Cox P M 2011 *Phil. Trans. R. Soc.* A **369** 868–86

[29] Milik A, Prskawetz A, Feichtinger G and Sanderson W C 1996 *Environ. Mod. Asse.* **1** 3–17

[30] Turchin P 2009 *Ann. NY Acad. Sci.* **1162** 1–17

[31] Anderies J M, Carpenter S R, Steffen W and Rockström J 2013 *Environ. Res. Lett.* **8** 044048

[32] Heck V, Donges J F and Lucht W 2016 *Earth Syst. Dyn.* **7** 783–96

[33] Forrester J W 1971 *World Dynamics* (Cambridge, MA: Wright-Allen Press)

[34] Millennium Ecosystem Assessment 2005 *Ecosystems and Human Well-being: Synthesis* (Washington, DC: Island Press)

[35] Haines-Young R H and Potschin M B 2010 The links between biodiversity, ecosystem services and human well-being *Ecosystems Ecology: A New Synthesis* ed D G Raffaelli and C L J Frid (Cambridge: Cambridge University Press)

[36] Ciais P *et al* 2013 Carbon and Other Biogeochemical Cycles *Climate Change 2013: The Physical Science Basis. Working Group I Contribution to the Fifth Assessment Report of the IPCC* 465–570

[37] Kremer M 1993 *Quart. J. Eco.* **108** 681–716

[38] Malthus T R 1798 *An essay on the principle of population, as it affects the future improvement of society* (London)

[39] Lotka A J 1910 *J. Phy. Chem.* **14** 271–74

[40] Usher D 1989 *Am. Econ. Rev.* **79** 1031–44

[41] Turchin P and Korotayev A V 2006 *Soc. Evol. His.* **5** 112–47

[42] Nefedov S 2014 *Soc. Evol. His.* **13** 172–84

[43] Kuznetsov Y A 1998 *Elements of Applied Bifurcation Theory* (New York: Springer)

[44] Steffen W, Broadgate W, Deutsch L, Gaffney O and Ludwig C 2015 *Anthropocene Rev.* **2** 1–18

[45] The World Bank 2016 World Development Indicators

[46] Nadiri M I and Prucha I 1996 *Econ. Inq.* **34** 43–56

[47] McKendry P 2002 *Bioresour. Technol.* **83** 37–46

[48] Krausmann F, Erb K H, Gingrich S, Haberl H, Bondeau A, Gaube V, Lauk C, Plutzar C and Searchinger T D 2013 *Proc. Natl Acad. Sci. USA* **110** 10324–9

[49] Clewley R 2012 *PLoS Comput. Biol.* **8** e1002628

Reproduced from: J. Nitzbon et al., Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., vol. 12, no. 7, 2017, doi: 10.1088/1748-9326/aa7581. Published under Creative Commons Attribution License 3.0 (CC BY).

2
Towards a unified analytical framework

In this second section, we present current approaches to model up-to-planetary-scale social-ecological dynamics and introduce a framework to unify the existing models.

The integration of human behaviour into formal Earth system models requires crucial assumptions about actors and their goals, behavioural options, and decision rules, as well as modelling decisions regarding human social interactions and the aggregation of individuals' behaviour.

In the first paper in this section, "Towards representing human behavior and decision-making in Earth system models" [Müller-Hansen et al., 2017b], we reviewed existing modeling approaches and techniques from various disciplines and found a very heterogeneous and diverse modeling landscape.

In order to structure future approaches, we proposed in "Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops" [Donges et al., 2018] three taxa for modelled subsystems: (i) biophysical, (ii) socio-cultural, and (iii) socio-metabolic. Furthermore, we introduced the model category of 'World-Earth models' (WEMs), i.e., models of social-ecological coevolution on up to planetary scales.

For the specific case of social tipping systems, we present in "Social tipping processes for sustainability: An analytical framework" [Winkelmann, R. and Donges, J. F. and Smith, E. K. and Milkoreit, M. et al., 2020] an analytical framework including a formal definition for social tipping processes and filtering criteria for those processes that could be decisive for future trajectories to global sustainability in the Anthropocene.

Building on these works, in "Earth system modeling with endogenous and dynamic human societies: the copan:CORE open World–Earth modeling framework" [Donges, J. F. and Heitzig, J. et al., 2020] we introduced design principles for constructing World-Earth models and presented an open-source software that teams of researchers with different backgrounds can use for implementing such models in a highly modular way, using a combination of equation-based and agent-based model components.

Earth Syst. Dynam., 8, 977–1007, 2017
https://doi.org/10.5194/esd-8-977-2017

Earth System
Dynamics
Open Access

Towards representing human behavior and decision making in Earth system models – an overview of techniques and approaches

Finn Müller-Hansen[1,2], Maja Schlüter[3], Michael Mäs[4], Jonathan F. Donges[1,3], Jakob J. Kolb[1,2], Kirsten Thonicke[1], and Jobst Heitzig[1]

[1]Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association,
P.O. Box 60 12 03, 14412 Potsdam, Germany
[2]Department of Physics, Humboldt University Berlin, Newtonstraße 15, 12489 Berlin, Germany
[3]Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden
[4]Department of Sociology and ICS, University of Groningen, Grote Rozenstraat 31, 9712 TG Groningen,
the Netherlands

Correspondence to: Finn Müller-Hansen (mhansen@pik-potsdam.de)

Received: 26 February 2017 – Discussion started: 2 March 2017
Accepted: 16 August 2017 – Published: 8 November 2017

Abstract. Today, humans have a critical impact on the Earth system and vice versa, which can generate complex feedback processes between social and ecological dynamics. Integrating human behavior into formal Earth system models (ESMs), however, requires crucial modeling assumptions about actors and their goals, behavioral options, and decision rules, as well as modeling decisions regarding human social interactions and the aggregation of individuals' behavior. Here, we review existing modeling approaches and techniques from various disciplines and schools of thought dealing with human behavior at different levels of decision making. We demonstrate modelers' often vast degrees of freedom but also seek to make modelers aware of the often crucial consequences of seemingly innocent modeling assumptions.

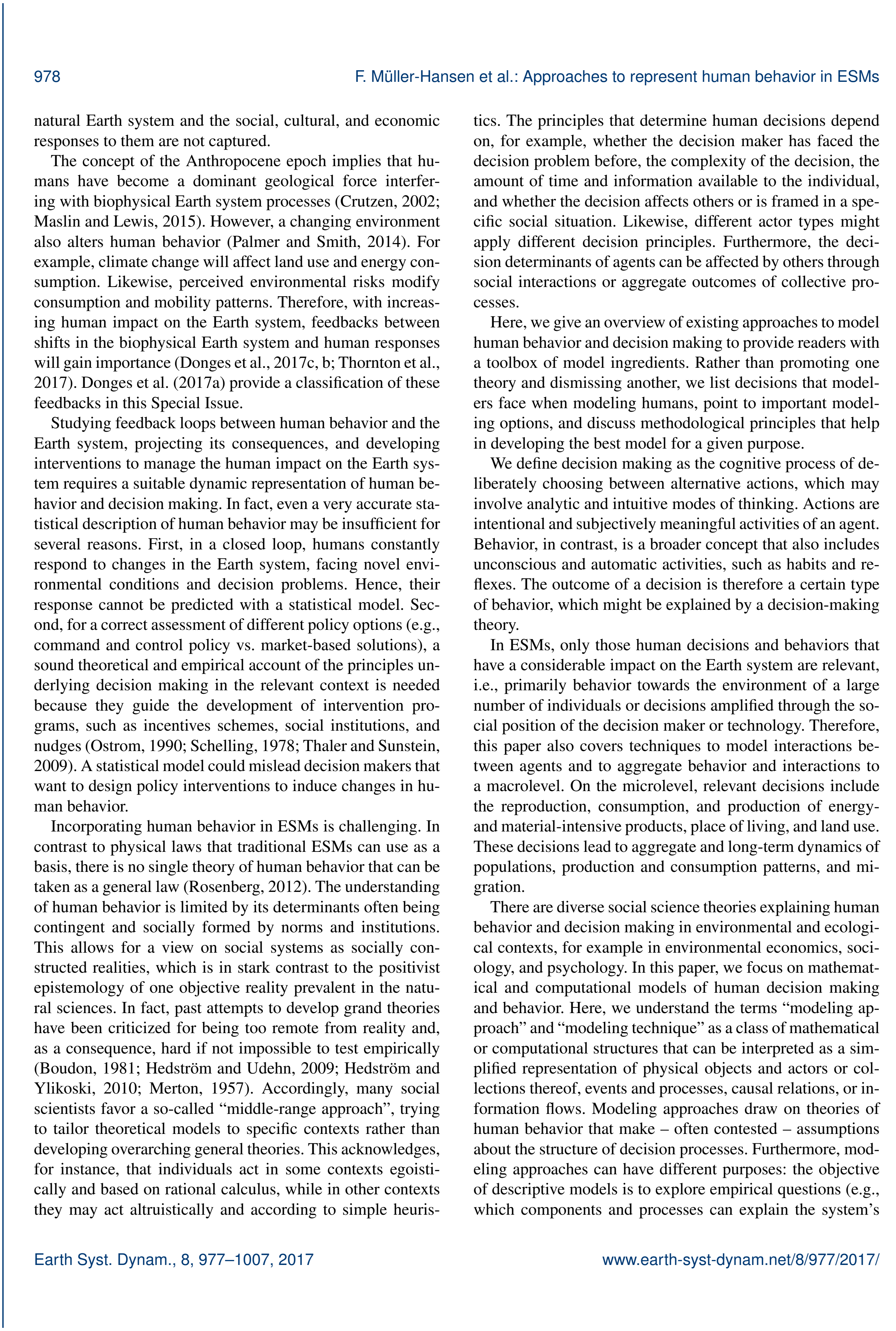
After discussing which socioeconomic units are potentially important for ESMs, we compare models of individual decision making that correspond to alternative behavioral theories and that make diverse modeling assumptions about individuals' preferences, beliefs, decision rules, and foresight. We review approaches to model social interaction, covering game theoretic frameworks, models of social influence, and network models. Finally, we discuss approaches to studying how the behavior of individuals, groups, and organizations can aggregate to complex collective phenomena, discussing agent-based, statistical, and representative-agent modeling and economic macro-dynamics. We illustrate the main ingredients of modeling techniques with examples from land-use dynamics as one of the main drivers of environmental change bridging local to global scales.

1 Introduction

Even though Earth system models (ESMs) are used to study human impacts on the complex interdependencies between various compartments of the Earth, humans are not represented explicitly in these models. ESMs usually consider human influence in terms of scenarios for comparison of the impacts of alternative narratives about the future development of key socioeconomic characteristics. For instance, the IPCC process uses integrated assessment models to compute plausible future emission pathways from energy and land use for different scenarios of climate mitigation. These projections determine the radiative forcing used as external input in ESMs to study its natural impacts (Moss et al., 2010; IPCC, 2014). The latter can, however, have socioeconomic consequences that may be fed back into the scenario process. However, the complex interplay of the dynamics of the

Published by Copernicus Publications on behalf of the European Geosciences Union.

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).



natural Earth system and the social, cultural, and economic responses to them are not captured.

The concept of the Anthropocene epoch implies that humans have become a dominant geological force interfering with biophysical Earth system processes (Crutzen, 2002; Maslin and Lewis, 2015). However, a changing environment also alters human behavior (Palmer and Smith, 2014). For example, climate change will affect land use and energy consumption. Likewise, perceived environmental risks modify consumption and mobility patterns. Therefore, with increasing human impact on the Earth system, feedbacks between shifts in the biophysical Earth system and human responses will gain importance (Donges et al., 2017c, b; Thornton et al., 2017). Donges et al. (2017a) provide a classification of these feedbacks in this Special Issue.

Studying feedback loops between human behavior and the Earth system, projecting its consequences, and developing interventions to manage the human impact on the Earth system requires a suitable dynamic representation of human behavior and decision making. In fact, even a very accurate statistical description of human behavior may be insufficient for several reasons. First, in a closed loop, humans constantly respond to changes in the Earth system, facing novel environmental conditions and decision problems. Hence, their response cannot be predicted with a statistical model. Second, for a correct assessment of different policy options (e.g., command and control policy vs. market-based solutions), a sound theoretical and empirical account of the principles underlying decision making in the relevant context is needed because they guide the development of intervention programs, such as incentives schemes, social institutions, and nudges (Ostrom, 1990; Schelling, 1978; Thaler and Sunstein, 2009). A statistical model could mislead decision makers that want to design policy interventions to induce changes in human behavior.

Incorporating human behavior in ESMs is challenging. In contrast to physical laws that traditional ESMs can use as a basis, there is no single theory of human behavior that can be taken as a general law (Rosenberg, 2012). The understanding of human behavior is limited by its determinants often being contingent and socially formed by norms and institutions. This allows for a view on social systems as socially constructed realities, which is in stark contrast to the positivist epistemology of one objective reality prevalent in the natural sciences. In fact, past attempts to develop grand theories have been criticized for being too remote from reality and, as a consequence, hard if not impossible to test empirically (Boudon, 1981; Hedström and Udehn, 2009; Hedström and Ylikoski, 2010; Merton, 1957). Accordingly, many social scientists favor a so-called "middle-range approach", trying to tailor theoretical models to specific contexts rather than developing overarching general theories. This acknowledges, for instance, that individuals act in some contexts egoistically and based on rational calculus, while in other contexts they may act altruistically and according to simple heuristics. The principles that determine human decisions depend on, for example, whether the decision maker has faced the decision problem before, the complexity of the decision, the amount of time and information available to the individual, and whether the decision affects others or is framed in a specific social situation. Likewise, different actor types might apply different decision principles. Furthermore, the decision determinants of agents can be affected by others through social interactions or aggregate outcomes of collective processes.

Here, we give an overview of existing approaches to model human behavior and decision making to provide readers with a toolbox of model ingredients. Rather than promoting one theory and dismissing another, we list decisions that modelers face when modeling humans, point to important modeling options, and discuss methodological principles that help in developing the best model for a given purpose.

We define decision making as the cognitive process of deliberately choosing between alternative actions, which may involve analytic and intuitive modes of thinking. Actions are intentional and subjectively meaningful activities of an agent. Behavior, in contrast, is a broader concept that also includes unconscious and automatic activities, such as habits and reflexes. The outcome of a decision is therefore a certain type of behavior, which might be explained by a decision-making theory.

In ESMs, only those human decisions and behaviors that have a considerable impact on the Earth system are relevant, i.e., primarily behavior towards the environment of a large number of individuals or decisions amplified through the social position of the decision maker or technology. Therefore, this paper also covers techniques to model interactions between agents and to aggregate behavior and interactions to a macrolevel. On the microlevel, relevant decisions include the reproduction, consumption, and production of energy- and material-intensive products, place of living, and land use. These decisions lead to aggregate and long-term dynamics of populations, production and consumption patterns, and migration.

There are diverse social science theories explaining human behavior and decision making in environmental and ecological contexts, for example in environmental economics, sociology, and psychology. In this paper, we focus on mathematical and computational models of human decision making and behavior. Here, we understand the terms "modeling approach" and "modeling technique" as a class of mathematical or computational structures that can be interpreted as a simplified representation of physical objects and actors or collections thereof, events and processes, causal relations, or information flows. Modeling approaches draw on theories of human behavior that make – often contested – assumptions about the structure of decision processes. Furthermore, modeling approaches can have different purposes: the objective of descriptive models is to explore empirical questions (e.g., which components and processes can explain the system's

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

dynamics), while normative models aim at answering ethical questions (e.g., which policy we should choose to reach a certain goal).

Recent reviews focus on existing modeling approaches and theories that are applied in the context of environmental management and change. For example, Verburg et al. (2016) assess existing modeling approaches and identify challenges for improving these models in order to better understand Anthropocene dynamics. An (2012), Meyfroidt (2013), and Schlüter et al. (2017) focus on cognitive and behavioral theories in ecological contexts, providing an overview for developers of agent-based, land-use, and social–ecological models. Cooke et al. (2009) and Balint et al. (2017) review different micro- and macro-approaches with applications to agroecology and the economics of climate change, respectively. The present paper complements this literature by reviewing modeling approaches of (1) individual agent behavior, (2) agent interactions, and (3) the aggregation of individual behaviors with the aim of supporting the integration of human decision making and behavior into Earth system models. The combination of these three different categories is crucial to describe human behavior at scales relevant for Earth system dynamics. Furthermore, this review highlights the strengths and limitations of different approaches by connecting the modeling techniques and their underlying assumptions about human behavior and discusses criteria to guide modeling choices.

Our survey of techniques has a bias towards economic modeling techniques for two simple reasons. First, economics is the social science discipline that has the longest and strongest tradition in the formal modeling of human decision making. Second, economics focuses on the study of production and consumption as well as the allocation of scarce resources. In most industrialized countries today, a major part of human interactions with the environment is mediated through markets, which are central in economic analyses. This review aims to go beyond the often narrow framing of economic approaches while at the same time not ignoring important economic insights. For instance, consumption and production decisions not only follow purely economic calculations, but are also deeply influenced, for instance, by behavioral patterns, traditions, and social norms (The World Bank, 2015).

Because we discuss different approaches to model decision making and behavior from various disciplinary or subdisciplinary scientific fields, there are considerable differences in terminology that make a harmonized presentation of the material challenging. For example, the same terms are used to describe quite separate varieties of an approach in different fields, and different terms from separate fields may refer to very similar approaches. We adopt a terminology that aims for a better interdisciplinary understanding and point out different understandings of contested terms where we are aware of them.

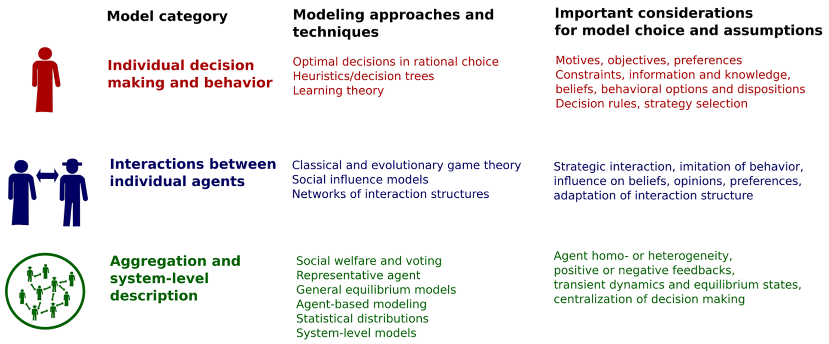
This paper works with land-use change as a guiding and illustrative example. Land-cover change and land use make up the second-largest source of greenhouse gases – besides the burning of fossil fuels – and thus contribute strongly to climate change. Behavioral responses related to land use will play a crucial role for successful mitigation and adaptation to projected climatic changes, thereby challenging modelers to represent decision making in models of land-use change (Brown et al., 2017). The complexity of land-use change provides various examples of how collective and individual decision making interacts with the environment across spatial scales and organizational levels. Land-use models consider environmental conditions as important factors in decision-making processes, giving rise to feedbacks between environmental and socioeconomic dynamics (Brown et al., 2016). However, this paper does not provide an exhaustive overview of existing land-use models. For this purpose, the reader is referred to the various reviews in the literature (e.g., Baker, 1989; Brown et al., 2004; Michetti, 2012; Groeneveld et al., 2017).

The remainder of the paper is organized as follows. In Sect. 2, we give an overview of different levels of description of social systems and the socioeconomic units or agents associated with them. Sections 3–5 form the main part of the paper, presenting different modeling techniques and their underlying assumptions about human decision making and behavior. First, Sect. 3 introduces approaches to model individual decisions and behavior from rational choice to learning theories. Many of these techniques can be used to also model higher-level social entities. Second, Sect. 4 puts the focus on techniques for modeling interactions between agents. Strategic interactions and social influence are significant determinants of individual decisions and therefore important for long-term changes in collective behavior, i.e., the group outcome of mutually dependent individual decisions. Third, Sect. 5 reviews different aggregation techniques that allow for a description of human activities at the level of social collectives or systems. These approaches make use of simplifications to scale up theories about individual decision making. Figure 1 summarizes these main parts of the paper, the corresponding modeling approaches, and important considerations for model selection, which we discuss in detail in Sect. 6. The discussion also reflects on important distinctions between models of natural and social systems that are crucial to consider when including human behavior into ESMs. The paper concludes with remarks on the remaining challenges for this endeavor.

2 The challenge: modeling decision making and behavior across different levels of organization

The decision making and behavior of humans can be described and analyzed at different levels of social systems. While decisions are made and behavior is performed by in-

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).



Model category	Modeling approaches and techniques	Important considerations for model choice and assumptions
Individual decision making and behavior	Optimal decisions in rational choice Heuristics/decision trees Learning theory	Motives, objectives, preferences Constraints, information and knowledge, beliefs, behavioral options and dispositions Decision rules, strategy selection
Interactions between individual agents	Classical and evolutionary game theory Social influence models Networks of interaction structures	Strategic interaction, imitation of behavior, influence on beliefs, opinions, preferences, adaptation of interaction structure
Aggregation and system-level description	Social welfare and voting Representative agent General equilibrium models Agent-based modeling Statistical distributions System-level models	Agent homo- or heterogeneity, positive or negative feedbacks, transient dynamics and equilibrium states, centralization of decision making

Figure 1. Overview of modeling categories, corresponding modeling approaches, and techniques discussed in this paper and important considerations for model choice and assumptions about human behavior and decision making.

dividual humans, it is often useful to not represent individual humans in a model but to treat social collectives, such as households, neighborhoods, cities, political and economic organizations, and states, as decision makers or agents.

Figure 2 shows a hierarchy of socioeconomic units, i.e., the groups, organizations, and structures of individuals that play a crucial role in human interactions with the Earth system. We consider a broad scheme of levels ranging from the microlevel across intermediate levels to the global level. This hierarchy of socioeconomic units is not only distinguishable by level of complexity but also by the different spatial scales involved. However, there is no one-to-one correspondence. For instance, some individuals have impacts at the global level, while many transnational organizations operate at specific local levels. Especially in the context of human–environment interactions in ESMs, scaling and spatial extent are therefore important issues (Gibson et al., 2000). Furthermore, we note that the strict separation between a microlevel and macrolevel may result in treating very different phenomena alike. For instance, many economic models describe both small businesses and transnational corporations as actors on the microlevel and model their decision processes with the same set of assumptions, even though they operate very differently.

One major challenge for modeling humans in the Earth system is therefore to bridge the diverse levels between individuals and the global scale, thereby integrating different levels of social organization and spatial and temporal scales.

The relation between individual agents and social collectives and structures has been the subject of considerable debate in the social sciences. In the social scientific tradition of methodological individualism[1], the analysis aims to explain social macro-phenomena, for example phenomena at the level of groups, organizations, or societies, with theories of individual behavior. This approach deviates from structuralist traditions, which claim that collective phenomena are of their own kind and thus cannot be traced back to the behavior of individuals (Durkheim, 2014). Positions between these two extremes emphasize the interdependency of individual agents and social structure, which is understood as an emerging phenomenon that stabilizes particular behaviors (Coleman, 1994; Homans, 1950). While it very much depends on the purpose of the given modeling exercise whether the model should represent individuals or collectives, we mainly focus here on the research tradition that acknowledges the fact that complex and unexpected collective phenomena can arise from the interplay of individual behavior.

In Table 1, we provide an overview of socioeconomic units at different levels that are potentially important for Earth system modeling. We list common theories, frameworks and assumptions made about decision making and behavior for these socioeconomic units and link them to scientific fields that focus on them.

At the microlevel, models consider individuals, households, families, and small businesses. For instance, individuals can make decisions as policy makers, investors, business managers, consumers, or resource users. At this level, decisions about lifestyle, consumption, individual natural resource use, migration, and reproduction are particularly relevant in the environmental context. Individual decisions have to be made by a large number of individuals or have to be reinforced by organizations, institutions, or technology to become relevant at the level of the Earth system. Individuals' participation in collective decision processes, such as voting, may also have consequences for the environment at a global level.

At various intermediate levels, communities and organizations like firms, political parties, labor unions, educational institutions, and nongovernmental and lobby organizations play a crucial role in shaping economic and political decisions and therefore have a huge impact on aggregate behavior. Governments at different levels representing different ter-

[1] We note that there are different accounts of methodological individualism, and it often remains unclear to what extent structural and interactionist elements can be part of an explanation (see Hodgson, 2007; Udehn, 2002).

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

Table 1. Overview of particular levels of description of socioeconomic units, associated scientific fields and communities, and some common approaches and assumptions about decisions and behavior. The list gives a broad overview but is far from being exhaustive.

Level	Socioeconomic units	Fields/communities	Common approaches and theories	Common assumptions about decision making
Micro	Individual humans	Psychology, neuroscience, sociology, economics, anthropology	Rational choice, bounded rationality, heuristics, learning theory, cognitive architectures	[All assumptions presented in this column]
	Households, families, small businesses	Economics, anthropology	Rational choice, heuristics, social influence	Maximization of consumption, leisure, profits
Intermediate	Communities (villages, neighborhoods), cities	Sociology, anthropology, urban studies	Social influence, networks	Transmission and evolution of cultural traits and traditions
	Political parties, NGOs, lobby organizations, educational institutions	Political science, sociology	Strategic decision making, public/social choice, social influence and evolutionary interactions	Agents form coalitions and cooperate to achieve goals, influenced by beliefs and opinions of others
	Governments	Political science, operations research	Strategic decision making, cost–benefit and welfare analysis, multi-criteria decision making,	Agents choose for the common good
	Nation states, societies	Economics, political science, sociology	welfare maximization, social choice	Majority vote
Global	Multinational firms, trade networks	Economics, management science	Rational choice	Maximization of profits or shareholder value
	Intergovernmental organizations	Political science (international relations)	Strategic decision making, cost–benefit analysis	Coalition formation

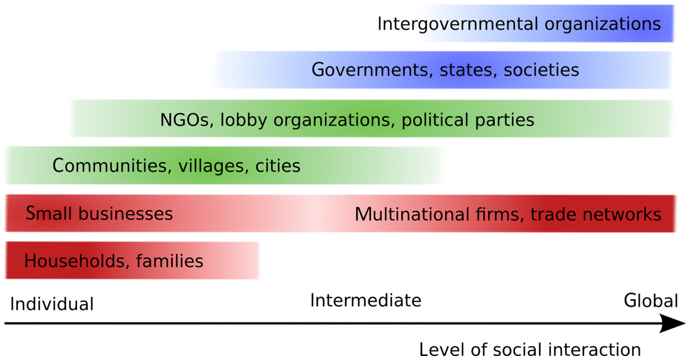

Intergovernmental organizations
Governments, states, societies
NGOs, lobby organizations, political parties
Communities, villages, cities
Small businesses
Multinational firms, trade networks
Households, families
Individual
Intermediate
Global
Level of social interaction

Figure 2. Socioeconomic units and their corresponding level and scales.

ritories, from cities to nation states, enact laws that strongly frame the economic and social activities of their citizens. Important decisions for the Earth system context include environmental regulations and standards, the production and distribution of commodities and assets, trade, the extraction and use of natural resources, and the development and building of physical infrastructures.

At the global level, multinational companies and intergovernmental organizations negotiate decisions. This level has considerable impacts on policy and business decisions even though it is remote from the daily life of most individuals. Often this level provides framing for activities on lower organizational levels and thus strongly influences the problem statements and perceived solutions, for instance regarding environmental issues. Decisions important for the Earth system at this level include international climate and trade agreements, the decisions of internationally operating corporations and financial institutions, and the adoption of global frameworks like the UN Sustainable Development Goals (United Nations General Assembly, 2015).

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

An overarching question that has triggered considerable debate between different disciplines is the allocation of agency at different levels of description. Even if individuals can decide between numerous options, the perception of options and decisions between them are shaped by social context and institutional embedding. Institutions[2] and organizations can display their own dynamics and lead to outcomes unintended by the individuals. On the other hand, social movements can initiate disruptive changes in institutional development. The attribution and perception of agency for a specific problem is therefore important for the choice of a suitable level of model description. The following section starts our discussion of different modeling techniques at the level of individual decision making and behavior.

3 Modeling individual behavior and decision making

In a nutshell, models of individual decision making and behavior differ with regard to their assumptions about three crucial determinants of human choices: goals, restrictions, and decision rules (Hedström, 2005; Lindenberg, 2001, 1990, 1985). First, the models assume that individuals have motives, goals, or preferences. That is, agents rank goods or outcomes in terms of their desirability and seek to realize highly ranked outcomes. A prominent but debated assumption of many models is that preferences or goals are assumed to be stable over time. Stable preferences are included to prevent researchers from developing trivial explanations, as a theory that models a given change in behavior only based on changed preferences does not have explanatory power. However, empirical research shows that preferences can change even in relatively short time frames (Ackermann et al., 2016). Changing individuals' goals or preferences is an important mechanism to affect their behavior, for example through policies, making flexible preferences particularly interesting for Earth system modelers.

Second, decision models make assumptions about restrictions and opportunities that constrain or help agents pursue their goals. For instance, each behavioral option comes with certain costs (e.g., money and time), and decision makers form more or less accurate beliefs about these costs and how likely they are to occur depending on the information available to the agent.

Third, models assume that agents apply some decision rule that translates their preferences and restrictions into a choice. Although decision rules differ very much in their complexity, they can be categorized into three types. First, there are decision rules that are forward looking. Rational choice theory, for instance, assumes that individuals list all positive and negative future consequences of a decision and choose the optimal option. Alternatively, backward-looking approaches, such as classical reinforcement learning, assume that actors remember the satisfaction experienced when they chose a given behavior in the past and tend to choose a behavior with a high satisfaction again. Finally, there are sideward-looking decision rules, which assume that actors adopt the behavior of others, for instance because they imitate successful others (Kandori et al., 1993). Theories assume different degrees of the context dependency of rules and make different implicit assumptions about the underlying cognitive capabilities of agents.

In the remainder of this section, we describe in more detail three important approaches to individual decision making and point out typical assumptions about motives, restrictions, and decision rules.

3.1 Optimal decisions and utility theory in rational choice models

Rational choice theory, a standard model in many social sciences (especially in economics) that is widely studied in mathematics, assumes that decision making is goal oriented: rational agents have preferences and choose the strategy with the expected outcome that is most preferred, given some external constraints and potentially based on their beliefs (represented by subjective probability distributions; see the beliefs, preferences, and constraints model in Gintis, 2009). It can either be used to represent actual behavior or serve as a normative benchmark for other theories of behavior.

How to judge the "rationality" of individual decisions is subject to ongoing debates. Opp (1999) distinguishes between strong rationality ("homo economicus"), assuming purely self-interested agents with unlimited cognitive capacities knowing all possible actions and probabilities of consequences, and weak rationality that makes less strong assumptions. Rabin (2002) distinguishes between standard and nonstandard assumptions regarding preferences, beliefs, and decision-making rules. Before discussing nonoptimal decision making in Sect. 3.2, we review here common assumptions on preferences and beliefs.

Usually, agents are assumed to be mainly self-interested, having fixed preferences regarding their personal consequences in possible futures and being indifferent to how a decision was made and to consequences for others. Exceptions are procedural (Hansson, 1996; Fehr and Schmidt, 1999) and other-regarding preferences (Mueller, 2003; Fehr and Fischbacher, 2003).

[2] The notion of institution is used in the literature with slightly different meanings: (1) formal and informal rules that shape behavior, (2) informal social order, i.e., regular patterns of behavior, and (3) organizations. Here, we adopt an understanding of institutions as formal (e.g., law, property rights) or informal rules (e.g., norms, religion). However, formal rules often manifest in social, political, and economic organizations and informal rules may be shaped by them.

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

Preferences can be modeled as binary preference relations, $x\ P_i\ y$, denoting that individual i prefers situation or outcome x to y. Most authors assume that P_i is complete (for every pair (x, y) either $x P_i y$ or $y P_i x$) and transitive (if $x P_i y$ and $y P_i z$ then $x P_i z$), which allows for the representation of the preferences with a utility function u_i (Von Neumann and Morgenstern, 1953).[3] Some authors also allow for incomplete or cyclic preferences (Fishburn, 1968; Heitzig and Simmons, 2012). In the land-use context, i could be a farmer, x might denote growing some traditional crops generating a moderate profit, and y growing hybrid seeds for more profit but making i dependent on the seed supplier. If i considers independence valuable enough to make up for the lower profit, $x\ P_i\ y$ would denote i's preference of x over y.

In decision making under uncertainty, agents have to choose between different risky prospects modeled as probability distributions $p(x)$ over outcomes x. In expected utility theory, p is preferred to p' if and only if $\sum_x p(x)u_i(x) > \sum_x p'(x)u_i(x)$. Empirical research shows that only a minority of people evaluate uncertainty in this risk-neutral way (Kahneman and Tversky, 1979). Prospect theory therefore models agents that overestimate small probabilities and evaluate outcomes relative to a reference point, which leads to risk-averse or risk-seeking behavior regarding losses or gains, respectively. (Kahneman and Tversky, 1979; Bruhin et al., 2010). A conceptual example from the land-use context illustrates decision making under risk. A farmer i might face the choice of whether to stick to her current crop x or switch to a new crop y. She may think that with 20 % probability the switch will result in a 50 % reduction in her profits, while with 80 % probability the profits would double. If her utility is proportional to the profits and she evaluates this uncertain prospect as described by expected utility theory, her gain from switching to y would be positive. If, however, she is averse to losses and thus conforms to prospect theory, she might evaluate the switch as negative and prefer to stick to x.

If several time points t are involved in a decision, agents are typically assumed to discount future consequences by using utility weights that decay in time and reflect the agent's time preferences. Discounted utility quantifies the present desirability of some utility obtained in the future. Most authors use exponentially decaying weights of the form e^{-rt} with a discounting rate $r > 0$ because this makes the evaluation independent of its time point. However, empirical studies suggest that people often use slower decaying weights (e.g., hyperbolic discounting), especially in the presence of uncertainty (Ainslie and Haslam, 1992; Jamison and Jamison, 2011), although this might lead to time-inconsistent choices that appear suboptimal at a later time. A farmer i may compare different crops not only by next year's expected profit $u_i(x, 1)$ but, due to the various crops' different effects on future soil quality, also by future years' profits $u_i(x, t)$ for $t > 1$. Crop y might promise higher yields than x in the short run but lower ones in the long run due to faster soil depletion. If i is "patient", having small r, she might prefer $y\ P_i\ x$ even though $u_i(x, 1) > u_i(y, 1)$.

Preferences can be aggregated not only in time but also across several interrelated issues or consequences. For example, consumer theory (Varian, 2010) models preferences over consumption bundles by combining the utility derived from consuming different products into a total consumption utility and simply adding up these utilities or combining them in some nonlinear way with imperfect substitutability of goods (Leontief, Cobb–Douglas, or CES utility functions). A farmers' utility from leisure time and crop yield $y(l)$ depending on working time l might, for example, be combined using the Cobb–Douglas utility function $u_i = y^\alpha(12 - l)^{1-\alpha}$ for some elasticity $\alpha \in (0, 1)$.

Complex optimization problems arising from rational choice theory can be solved by mathematical programming, calculus of variations, and similar methods (see, e.g., Kamien and Schwartz, 2012; Chong and Zak, 2013). Optimal decisions under constraints are not only discussed as a description of human behavior, but are also often taken as the normative benchmark for comparison with other nonoptimal approaches that we discuss in Sect. 3.2.

Regarding decision modeling in ESMs, rational choice theory is useful when agents have clear goals and possess enough information and cognitive resources to assess the optimality of strategies. For instance, individuals' decisions regarding long-term investments or the decisions of organizations, such as firms or governments, in competitive situations can often be assumed to follow a rational choice model reasonably well. It can also be useful when actors make repeated similar decisions and can learn optimal strategies from fast feedback, making them behave "as if" they were rational.

3.2 Bounded rationality and heuristic decision making

Empirical research on human decision making finds that individual behavior depends on the framing and context of the decision (Tversky and Kahneman, 1974). Human decision making is characterized by deviations from the normative standards of the rational choice model, so-called cognitive biases, challenging the assumption that rational choice theory serves not only as a normative benchmark, but also as a descriptive model of individual decision making. Biases can be the result of time-limited information processing (Hilbert, 2012), heuristic decision making (Simon, 1956), or emotional influences (e.g., wishful thinking, Babad and Katz, 1991; Loewenstein and Lerner, 2003). Bounded rationality theory assumes that human decision making is constrained by the cognitive capabilities of the agents in addition to the constraints imposed by the environment and the available information about it (Simon, 1956, 1997). In the economic literature, non-transitive preferences, time-inconsistent discounting, and deviations from expected utility that we al-

[3] $u_i(x) > u_i(y)$ implies $x\ P_i\ y$, where u_i is only defined up to positive linear (affine) transformations.

ready introduced in the previous subsection are also often considered as boundedly rational (Gintis, 2009). Boundedly rational agents can be considered as *satisficers* that try to find a satisfying action in a situation given their available information and cognitive capabilities (Gigerenzer and Selten, 2002).

Constraints on information processing imply that agents do not integrate all the available information to compute the utility of every possible option in complex decision situations and choose an action with maximal utility. Instead, agents use heuristics to judge the available information and choose actions that lead to the more preferred outcome over less preferred ones. Gigerenzer and Gaissmaier (2011) define heuristics in decision making as a "strategy that ignores part of the information, with the goal of making decisions more quickly, frugally, and/or accurately than more complex methods." It is argued that instead of an all-purpose tool, the mind carries an "adaptive toolbox" of different heuristic decision schemes applicable in particular environments (Gigerenzer and Selten, 2002; Todd and Gigerenzer, 2007).

In general, heuristic rules are formalized either as decision trees or flowcharts and consist of three building blocks: one for information search, one for stopping the information search, and one to derive a decision from the information found. They evaluate a number of pieces of information – so-called cues – to either categorize a certain object or to choose between several options. Many heuristics evaluate these cues in a certain order and make a decision as soon as a cue value allows for classification or discriminates between options.

This is illustrated by means of the take-the-best heuristic: pieces of information (cues) are compared between alternatives according to a prescribed order, which is crucial for the decision process. At each step in the cue order, some information is searched for and evaluated. If the information does not allow for discrimination between the options, the process moves on to the next cue. This repeats as the process moves down the cue order until a cue is reached for which the differentiation between options is possible and the option with the higher cue value is chosen. Another notable example is the satisficing heuristic that evaluates information sequentially and chooses the first option satisfying certain criteria. Heuristics, especially cue orders, can be interpreted as encoding norms and preferences in individual decision making as they prioritize features of different options over others and hierarchically structure the evaluation of available information. An overview and explanation of numerous other decision heuristics can be found in the recent review paper by Gigerenzer and Gaissmaier (2011).

Gigerenzer and Todd (1999) question the usefulness of rational choice theory as the normative benchmark because it is not designed for so-called "large worlds" where information relevant for the decision process is either unknown or has to be estimated from small samples. Instead, they want to relieve heuristic decision making of its stigma of cognitive laziness, bias, and irrationality. With their account of ecological rationality, they suggest that heuristics can also serve as a normative choice model providing context-specific rules for normative questions. This is motivated by the observation that in many real-world situations, especially when high uncertainties are involved, some decision heuristics perform equally good or even better than more elaborated decision strategies (Dhami and Ayton, 2001; Dhami and Harries, 2001; Keller et al., 2014).

So far, heuristics have been used to describe decisions, for instance in consumer choice (Hauser et al., 2009), voter behavior (Lau and Redlawsk, 2006), and organizational behavior (Loock and Hinnen, 2015; Simon, 1997). However, fast and frugal decision heuristics are not yet commonly applied in dynamic modeling of human–nature interactions. One exception is the description of farmer and pastoralist behavior in a study of the origins of conflict in East Africa (Kennedy and Bassett, 2011). However, as the following example shows, similar decision trees have been used to model decision making in agent-based simulations of land-use change. The model by Deadman et al. (2004) describes colonist household decisions in the Amazon rainforest. Each household is a potential farmer who first checks whether a subsistence requirement is met. If this is not the case, the household farms annual crops. If the subsistence requirement is met, the household eventually plants perennials or breeds livestock depending on the soil quality. The model shows how heuristic decision trees can be used to simplify complex decision processes and represent them in an intelligible way. However, the example also shows the many degrees of freedom in the construction of heuristics, pointing at the difficulty to obtain these structures from empirical research.

Heuristics are a promising tool for including individual human decision making into ESMs because they can capture crucial choices in a computationally efficient way. In order to describe the long-term evolution of preferences, norms, and values relevant for human interactions with the Earth system, heuristics could also be used to model meta-decisions of preference or value adoption. Recent findings suggest that cue orders can spread via social learning and social influence (Gigerenzer et al., 2008; Hertwig and Herzog, 2009) analogously to norm and opinion spreading in social networks (see Sects. 4.3 and 4.4), which could be a promising approach to model social change. However, in contrast to fully rational decision making, it can be very challenging to aggregate heuristic decision making analytically to higher organizational levels. Therefore, approaches like agent-based modeling are suitable to explore the aggregate outcomes of many agents with such decision rules (see Sect. 5.5).

3.3 Learning theory

The approaches discussed in the previous two subsections mainly took the perspective of a forward-looking agent. Rational or boundedly rational actors optimize future payoffs based on information or beliefs about how their behavior af-

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

fects future payoffs, while the procedures to optimize may be more or less bounded. However, these techniques do not specify how the information is acquired and how the beliefs are formed. Computational learning theory focuses on behavior from a backward-looking perspective: an agent learned in the past that a certain action gives a reward that feels good or is satisfying and is therefore more likely to repeat this behavior. It can describe the adaptivity of agent behavior to a changing environment and is particularly suited for modeling behavior under limited information. To model the learning of agents, unsupervised learning techniques are mostly used because they do not require training with an external correction.

Reinforcement learning is such a technique that models how an agent maps environmental conditions to desirable actions in a way that optimizes a stream of rewards (and/or punishments). The obtained reward depends on the state of the environment and the chosen action, but may also be influenced by chosen actions and environmental conditions in the past. According to Macy et al. (2013), reinforcement learning differs from forward-looking behavioral models regarding three key aspects. (1) Because agents explore the likely consequences and learn from outcomes that actually occurred rather than those which are intended to occur but may only be obtained with a certain probability, reinforcement learning does not need to assume that the consequences are intended. (2) Decisions are guided by rewards or punishments that lead to approach or avoidance rather than by static utilities. (3) Learning is characterized by stepwise melioration and models the dynamic search for an optimum rather than assuming that the optimal strategy can be determined right away.

The learning process is modeled via a learning algorithm (e.g., Q-learning, SARSA learning, actor-critic learning) based on iteratively evaluating the current value of the environmental state utilizing a temporal difference error of expected value and experience value (Sutton and Barto, 1998). Artificial neural network algorithms can explore very high dimensional state and action spaces. Genetic algorithms, which are inspired by evolutionary mechanisms such as mutation and selection, are also applied to learning problems. The learning algorithm has to balance a trade-off between the exploration of actions with unknown consequences and the exploitation of current knowledge. In order to not exploit only the currently learned strategy, many algorithms use randomness to induce deviations from already learned behavior.

The environment in reinforcement learning problems is often modeled with Markovian transition probabilities. The special case of a single agent is called a Markov decision process (Bellman, 1957). In each of the discrete states of the environment the agent can choose from a set of possible actions. The choice then influences the transition probabilities to the next state and the reward. As an illustration, consider a farmer adapting her planting and irrigation practices to new climatic conditions. The environment could be modeled by a Markov process with different states of soil fertility and moisture, in which transitions between states reflect the influence of stochastic weather events. Without the possibility to acquire knowledge through other channels, she would explore different possible actions and evaluate how they change the yield (her reward). Eventually, through a trial-and-error process, her yield would increase on average.

A common approach to model the acquisition of subjective probabilities associated with the consequences of actions is Bayesian learning, which has also been applied to reinforcement learning problems (Vlassis et al., 2012). Starting with some prior probability (e.g., from some high-entropy "uninformative" distribution) $P(h_i)$ that some hypothesis h_i about the relation of actions and outcomes is true, new information or evidence $P(E)$ is used to update the subjective probability with the posterior $P(E|h_i)$ calculated with Bayes' theorem: $P(h_i|E) = P(E|h_i)P(h_i)/P(E)$ (Puga et al., 2015). The most probable hypothesis can then be chosen to determine further action.

By combining various approaches to model the acquisition of beliefs through learning, the formation of preferences and different decision rules discussed in the previous sections with further insights from psychology and neuroscience has led to the development of very diverse and detailed behavioral theories which are often formalized in cognitive architectures (Balke and Gilbert, 2014). These approaches can be used to describe human behavior in computational models, but are too complex and diverse to discuss them here in detail.

Learning and related theories that emphasize the adaptability of human behavior might be important building blocks to model the long-term evolution of human interactions with the Earth system from an individual perspective. On the other hand, they can capture short-term responses to drastically changing natural environments that are relevant, for instance, in the context of tipping elements in the Earth system.

Table 2 summarizes the approaches that focus on individual human behavior. Besides the forward- and backward-looking behavior that we introduced in this section, agents may exhibit sideways-looking behavior: agents can copy the behavior of successful others, thereby contributing to a social learning process. For this kind of behavior, interactions between different agents are crucial. This will be the focus of the next section.

4 Modeling interactions between agents

In the previous section, we discussed modeling approaches that focus on the choices of individuals that are confronted with a decision in a specified situation. In contrast, this section reviews techniques to model how actors interact with each other and influence or respond to each other's decisions. Interactions at the system level that are also aggrega-

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

Table 2. Summary table for individual behavior and decision making.

Theories	Key considerations	Strengths	Limitations
Optimal decisions in rational choice: individuals make the decision that maximizes their expected utility given economic, social, and environmental constraints.	What are the agent's preferences? What information (and beliefs) do they have?	Highly researched theory with strong theoretical foundation and many applications	Individuals assumed to have strong capabilities for information processing and perfect self-control
Bounded rationality and heuristic decision making: individuals have biases and heuristic decision rules that help them navigate complex environments effectively.	Which cue order is used to gather and evaluate information? When do agents stop gathering more information and decide?	Simple decision processes that capture observed biases in decision making	Suitable decision rules highly context dependent
Learning: agents explore possible actions through repeated learning from past experience.	How do agents interact with their environment? What is the trade-off between exploitation of knowledge and exploration of new options?	Captures information and belief acquisition processes	High degree of randomness in behavioral changes

tion mechanisms (e.g., voting procedures and markets) will be discussed in Sect. 5.

The section starts with a review of strategic interactions as modeled in classical game theory and dynamic interactions in evolutionary approaches. Then, we address models of social influence that are used to study opinion and preference formation or the transmission of cultural traits, i.e., culturally significant behaviors. Finally, we discuss how interaction structures can be modeled as dynamic networks.

4.1 Strategic interactions between rational agents: classical game theory

Game theory focuses on decision problems of "strategic interdependence", in which the utility that a decision maker (called the player) gets depends not only on her own decision, but also on the choices of others. These are often situations of conflict or cooperation. Players choose an action (behavioral option, control) based on a strategy, i.e., a rule specifying which action to take in a given situation. Classical game theory explores how rational actors identify strategies, usually assuming the rationality of other players. However, rational players can also base their choices on beliefs about others players' decisions, which can lead to an infinite regress of mutual beliefs about each other's decisions.

Formally, a game is described by what game theorists call a game form or mechanism. The game form specifies the actions $a_i(t)$ that agents can choose at well-defined time points t from an action set $A_i(t)$ that may vary over time, having to respect all kinds of situation-dependent rules. The game form may furthermore allow for communication with the other agent(s) (signaling) or binding agreements (commitment power). Simple social situations are formalized in so-called normal-form games represented by a payoff matrix specifying the individual utilities[4] for all possible action combinations, while more complex situations are modeled as a stepwise movement through the nodes of a decision tree or game tree (Gintis, 2009).

Classical game theory assumes that players form consistent beliefs about each other's unobservable strategies, in particular that the other's behavior results from an optimal strategy. However, multiplayer interaction and optimization often leads to recursive relationships between beliefs and strategies, which makes solving complex classical games often very difficult. Many problems have several solutions, called equilibria (not to be confused with the steady-state meaning of the word), and call for sophisticated nonlinear fixed-point solvers (Harsanyi and Selten, 1988). Only in special cases, for example in which players have complete information and moves are not simultaneous but alternating, game-theoretic equilibria can easily be predicted by simple solution concepts such as backwards induction (Gintis, 2009). In other cases, one can identify strategies and belief combinations consistent with the following two assumptions. First, each player eventually chooses a strategy that is optimal given her beliefs about all other players' strategies (rational behavior). Second, each player's eventual beliefs about other players' strategies are correct (rational expectations). The solutions are called Nash equilibria. However, many games have multiple Nash equilibria, and the question of which equilibrium will be selected arises.

[4] Note that despite the term "payoff matrix", these utilities are unexplained attributes of the agents and need not have a relation to monetary quantities.

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

Therefore, game theorists try to narrow down the likely strategy combinations by assuming additional forms of consistency and rationality (Aumann, 2006), such as consistency over time (sequential and subgame perfect equilibria), stability against small deviations (stable equilibria, Foster and Young, 1990), or small random mistakes (trembling hand perfect equilibria, Harsanyi and Selten, 1988). After a plausible strategic equilibrium has been identified, it can be used in a simulation of the actual behavior resulting from these strategies over time, possibly including noise and mistakes.

As an example from the land-use context, consider two farmers living on the same road. They get their irrigation water from the same stream. A dispute over the use of water emerges. Both may react to the actions of the other in several turns. The upstream farmer located at the end of the road may increase or decrease her water use and/or pay compensation for using too much water to the other. The downstream farmer at the entrance of the road may demand compensation or block the road and thereby cut the access of the upstream farmer to other supplies. A complex game tree encodes which actions are feasible at which moment and what are the consequences on players' utilities. If it is possible to specify the information and options available to the players at each time point, then a classical game theoretical analysis allows for the determination of the rational equilibrium strategies that the farmers would follow.

Classical game theory is widely applied to interactions in market settings in economics (see also Sect. 5.2), but increasingly also in the social and political sciences to political and voting behavior in public and social choice theory (see, e.g., Ordeshook, 1986; Mueller, 2003, and Sect. 5.1). For example, public choice theory studies strategic interactions between groups of politicians, bureaucrats, and voters with potentially completely different preferences and action sets.

While many simple models of strategic interactions between rational and selfish agents will predict only low levels of cooperation, more complex models can well explain how bilateral and multilateral cooperation, consensus, and stable social structure emerges (Kurths et al., 2015). This has been shown in contexts such as multiplayer public goods problems and international climate policy (e.g., Heitzig et al., 2011; Heitzig, 2013).

To model relevant decision processes in the Earth system, classical game-theoretic analysis could be used for describing strategic interactions between agents that could be assumed as highly rational and well informed, i.e., international negotiations of climate agreements between governments, bargaining between social partners, or monopolistic competition between firms. Similarly, international negotiations and their interactions with domestic policy can also be framed as two-level or multilevel games (as in some models of political science, e.g., Putnam, 1988; Lisowski, 2002). Furthermore, social choice theory could be used to simulate simple voting procedures that (to a certain extent) determine the goals of regional or national governments.

4.2 Interactions with dynamic strategies: evolutionary approaches and learning in game theory

In game-theoretic settings, complex individual behavioral rules are typically modeled as strategies specifying an action for each node in the game tree. Consider as an example the repeated version of the prisoners' dilemma in which each of two players can either "cooperate" or "defect" in each period (Aumann, 2006). A typical complex strategy in this game could involve reciprocity (defect temporarily after a defection of your opponent), forgiveness (every so often not reciprocate), and making up (do not defect again after being punished by a defection of your opponent after your own defection).

Many or even most nodes of a game tree will not be visited in the eventual realization of the game, and strategies may involve the deliberate randomization of actions. Therefore, strategies, unlike actual behavior, are principally unobservable, and assumptions about them are hard to validate. For this and other reasons, several kinds of additional assumptions are often made that constrain the set of strategies further that a player can choose, e.g., assuming only very short memory or low farsightedness (myopic behavior) and disallowing randomization, or allowing only strategies of a specific formal structure such as heuristics (see Sect. 3.2).

The water conflict example from Sect. 4.1 bears some similarity to the repeated prisoners' dilemma in that the farmers' possible actions can be interpreted as either defective (using too much water, blocking the road) or cooperative (not doing any of this, compensating for past defections). Assuming different levels of farsightedness may thus lead to radically different actions because myopic players would much more likely get trapped in a cycle of alternating defections than farsighted players. The latter would recognize some degree of forgiveness because that maximizes long-term payoff and would thus desist from defection with some probability. In any case, both farmers' choices can be modeled as depending on what they believe the other will likely do or how she will react to the last action.

Evolutionary approaches in game theory study the interaction of different strategies and analyze which strategies prevail on a population level as a result of selection mechanisms. Thus, in contrast to classical game theory, evolutionary approaches focus on the dynamics of strategy selection in populations. The agent's strategies may be hardwired, acquired, or adapted by learning (Fudenberg and Levine, 1998; Macy and Flache, 2002). Although many evolutionary techniques in game theory are used in biology to study biological evolution (variation through mutation, selection by fitness, and reproduction with inheritance), evolutionary game theory can be used to study all kinds of strategy changes in game-theoretic settings, for instance cultural evolution (transmission of memes), social learning through the imitation of successful strategies, or the emergence of cooperation (Axelrod, 1984, 1997).

In an evolutionary game, a population of agents is divided into factions with different strategies. They interact in a formal game (given by a payoff matrix or game tree, see Sect. 4.1), in which their strategy results in a fitness (or payoff). The factions change according to some replicator rules that depend on the acquired fitness. This can be modeled using different techniques. Simple evolutionary games in well-mixed large populations can be described with replicator equations. The dynamics describing the relative change in the factions with a particular strategy is proportional to the deviation of the fitness of this faction from the average fitness (Nowak, 2006).

Alternatively, the behavior resulting from evolutionary interactions is often easy to simulate numerically as a discrete-time dynamical system even for large numbers of players if the individual action sets are finite or low-dimensional and only certain simple types of strategies are considered. This type of agent-based model (see Sect. 5.5) simply implements features such as mutation or experimentation and replication via strategy transfer (e.g., imitation and inheritance) at the microlevel. Combined with network approaches (see Sect. 4.4), the influence of interaction structure can also be studied (Szabó and Fáth, 2007; Perc and Szolnoki, 2010). Strategies can be characterized as evolutionary stable if a population with this strategy cannot be invaded by another, initially rare strategy. If a strategy is furthermore stable for finite populations or noisy dynamics, it is called stochastically stable.

In our water conflict example, the farmers could use a heuristic strategy (see Sect. 3.2) that determines how much water they extract given the actions of the other. The evolution of the strategies could either be modeled with a learning algorithm, repeating the game again and again. Alternatively, to determine feasible strategies in an evolutionary setting, a meta-model could consider an ensemble of similar villages consisting of two farmers. The strategies of the farmers would then be the result of either an imitation process between the villages or of an evolutionary process, assuming that less successful villages die out over time.

Evolutionary approaches to game theory are a promising framework to better understand the prevalence of certain human behaviors regarding interaction with the Earth system. This is especially interesting regarding the modeling of long-term cultural evolution and changes in individuals' goals, beliefs, and decision strategies or the transmission of endogenous preferences (Bowles, 1998).

4.3 Modeling social influence

Human behavior and its determinants (beliefs, goals, and preferences) are strongly shaped by social influence, which can result from various cognitive processes. Individuals may be convinced by persuasive arguments (Myers, 1982), aim to be similar to esteemed others (Akers et al., 1979), be unsure about what is the best behavior in a given situation (Bikhchandani et al., 1992), or perceive social pressure to conform with others (Wood, 2000; Festinger et al., 1950; Homans, 1950).

Models of social influence allow for the study of the outcomes of repeated influence in social networks and have been used to explain the formation of consensus, the development of monoculture, the emergence of clustered opinion distributions, and the emergence of opinion polarization, for instance. Models of social influence are very general and can be applied to any setting in which individuals exert some form of influence on each other. However, seemingly innocent differences in the formal implementation of social influence can have decisive effects on the model outcomes, as the following list of important modeling decisions documents.

A first question is how social influence changes individual attributes. For example, a farmer deciding when to till his field might either choose the date that most of his neighbors think is best, take the average of the proposed dates, or even try to counter coordinate with disliked farmers. Classical models incorporate influence as averaging, which implies that interacting individuals always grow more similar over time (Friedkin and Johnsen, 2011). Averaging is an accepted and empirically supported model of influence resulting, for instance, from the social pressure that an actor exerts on someone else (Takács et al., 2016). Models assume different forms of averaging. Rather than following the arithmetic average of all opinions, actors might only consider the majority view (Nowak et al., 1990). In other models, social influence can lead to polarization (Myers, 1982). For instance, in models of argument communication, actor opinions can turn more extreme when the interaction partners provide them with new arguments that support their own opinion (Mäs and Flache, 2013; Mäs et al., 2013).

Second, modelers need to decide whether there is just one or multiple dimensions of influence. For instance, it is often argued that political opinions are multidimensional and cannot be captured by the one-dimensional left–right spectrum. Explaining the dynamics of opinion polarization and clustering is often more difficult when multiple dimensions are taken into account (Axelrod, 1997). Additionally, model predictions often depend on whether the influence dimension is a discrete or a continuous variable. Models of individuals' decisions about certain policies often model the decisions as binary choices (Sznajd-Weron and Sznajd, 2000; Martins, 2008). However, binary scales fail to capture the fact that many opinions vary on a continuous scale and that differences between individuals can therefore also increase in a single dimension (Feldman, 2011; Jones, 2002; Stroud, 2010). Therefore, models that describe opinion polarization usually treat opinions as continuous attributes.

A third critical question is how the interaction process is modeled. In models of opinion dynamics, for example, influ-

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

ence is bidirectional in that an actor who exerts influence on someone else can also be influenced by the other (Macy et al., 2013; Mäs et al., 2010). In diffusion models, in contrast, the effective influence is directed. For instance, information can spread only from informed to uninformed individuals, but not the other way around. Furthermore, actors may be influenced dyadically or multilaterally. Model outcomes often depend on whether the influence that a group exerts on an actor is modeled as a sequence of events involving dyads of actors or as a single opinion update in which the actor considers all contacts' influences at once (Flache and Macy, 2011; Lorenz, 2005; Huckfeldt et al., 2004). In models that assume binary influence dimensions, for instance, dyadic influence implies that an agent copies a trait from her interaction partner. When influence is multilateral, agents aggregate the influence exerted by multiple interaction partners (using, e.g., the mode of the neighbors' opinions), which can imply that agents with rare traits are not considered even though they would have an influence in the case of dyadic influence events. For example, a farmer seeking advice on whether to adopt a new technology can either consult his friends one after another or all together, likely leading to different outcomes if they have different opinions on the matter.

Fourth, agents may slightly deviate from the influence of their contacts. The exact type of these deviations affects model outcomes and can introduce a source of diversity into models of social influence (Mäs et al., 2010; Pineda et al., 2009; Kurahashi-Nakamura et al., 2016). For instance, some models of continuous opinion dynamics include deviations as Gaussian noise, i.e., random values drawn from a normal distribution. In such a model, opinions in homogeneous subgroups will fluctuate randomly and subgroups with similar opinions can merge that would have remained split in a model without deviations (Mäs et al., 2010). In other contexts, deviations are better modeled by uniformly distributed noise, assuming that big deviations are as likely as small ones. This can help to explain, for instance, the emergence and stability of subgroups with different opinions that do not emerge in settings with Gaussian noise[5] (Pineda et al., 2009).

Finally, the effects of social influence depend on the structure of the network that determines who influences whom. Complex dynamics can arise when this interaction network is dynamic and depends on the attributes of the agents, as we discuss in the following section.

Models of social influence are a promising approach to explore how social transitions interact with the Earth system, for example transitions of norms regarding admissible resource use and emissions, lifestyle changes, and adoption of new technology. They can be used to explore the conditions under which social learning enables groups of agents to adopt sustainable management practices.

4.4 Modeling the interaction structure: (adaptive) network approaches

In most of the models discussed in the previous section, the social network is formally modeled as a graph (the mathematical notion for a network): a collection of nodes that are connected by links. In this mathematical framework, nodes (vertices) represent agents and links (edges) indicate interaction, communication, or a social relationship. Agents can only interact and thus influence each other if they are connected by a link in the underlying network.

Classical social influence models study the dynamics of influence on static networks, assuming that agents are always affected by the same subset of interaction partners (e.g., DeGroot, 1974; French, 1956; Friedkin and Johnsen, 2011). These networks can be undirected or directed, possibly restricting the direction of influence, but their structure does not change over time. Furthermore, the topology of the network, i.e., the arrangement of links, can be more or less random or regular, clustered, and hierarchical. In social influence models on static networks, connected populations will usually reach consensus in the long run.

Especially when modeling social processes over longer timescales, it is reasonable to assume that the social network is dynamic, i.e., that its structure evolves over time. This time evolution can be independent of the dynamics on the network and encoded in a temporal network (Holme and Saramäki, 2012). However, for many social processes, the structure of the social network and the dynamics on the network (e.g., social influence) interact. Adaptive network models make the removal of existing and the formation of new links between agents dependent on attributes of the agents by building on the insight that the social structure influences the behavior, opinions, or beliefs of individual actors, which in turn drives changes in social structure (Gross and Blasius, 2008).

Local update rules for the social network structure and the agent behavior can be chosen very flexibly. Changes in agent behaviors may be governed by rules such as random or boundedly rational imitation of the behavior of network neighbors (see above). Update rules for the network structure are often based on the insight that agents tend to be influenced by similar others and ignore those who hold too-distant views (Wimmer and Lewis, 2010; McPherson et al., 2001; Lazarsfeld and Merton, 1954). Many models assume that agents with similar characteristics tend to form new links between each other (homophily) while breaking links with agents having diverging characteristics (Axelrod, 1997; Hegselmann and Krause, 2002; Deffuant et al., 2005). In adaptive network models, homophily in combination with social influence generates a positive feedback loop: influence increases similarity, which leads to more influence and so on. Such models can explain, for instance, the emergence and stability of multiple internally homogeneous but mutually different subgroups. Other applications of coevolutionary network models allow us to understand the presence

[5] Gaussian noise needs to be very strong to generate enough diversity for the emergence of subgroups with different opinions. However, when noise is strong, subgroups will not be stable.

of social tipping points in opinion formation (Holme and Newman, 2006), epidemic spreading (Gross et al., 2006), the emergence of cooperation in social dilemmas (Perc and Szolnoki, 2010), and the interdependence of coalition formation with social networks (Auer et al., 2015). Such adaptive network models exhibit complex and nonlinear dynamics such as phase transitions (Holme and Newman, 2006), multi-stability (Wiedermann et al., 2015), oscillations in both agent states and network structure (Gross et al., 2006), and structural changes in network properties (Schleussner et al., 2016).

While adaptive networks have so far mostly been applied to networks of agents representing individuals, the framework can in principle be used to model coevolutionary dynamics on various levels of social interaction as introduced in Table 1. For instance, global complex network structures such as financial risk networks between banks, trade networks between countries, transportation networks between cities and other communication, organizational, and infrastructure networks can be modeled (Currarini et al., 2016). Furthermore, approaches such as multi-layer and hierarchical networks or networks of networks allow for the modeling of the interactions between different levels of a system (Boccaletti et al., 2014).

As an illustration, consider a community of agents each harvesting a renewable resource, for example wood from a forest. The agents interact on a social network, imitating the harvesting effort of neighbors that harvest more and may drop links to neighbors that use another effort. The interaction of the resource dynamics with the network dynamics either leads to a convergence of harvest efforts or a segregation of the community into groups with higher or lower effort depending on the model parameters (Wiedermann et al., 2015; Barfuss et al., 2017).

In the context of long timescales in the Earth system, the time evolution of social structures that determine interactions with the environment are particularly important. Adaptive networks offer a promising approach to modeling the structural change of the internal connectivity of a complex system (Lade et al., 2017). For example, this could be applied to explore mechanisms behind transitions between centralized and decentralized infrastructure and organizational networks.

Table 3 summarizes the different modeling approaches that focus on agent interactions in human decision making and behavior. These interactions occur between two or several agents. For including the effect of these interactions into ESMs, their aggregate effects need to be taken into account as well. Therefore, we introduce in the next section approaches that allow us to aggregate individual behavior and local interactions and to study the resulting macrolevel dynamics.

5 Aggregating behavior and decision making and modeling dynamics at the system level

So far, we focused on theories and modeling techniques that describe the decision processes and behavior of single actors, their interactions, and the interaction structure. This section builds on the previously discussed approaches and highlights different aggregation methods for the behavior of an ensemble or group of agents. This is an important step if models shall describe system-level outcomes or collective decision making and behavior in the context of Earth system modeling. Aggregation techniques link modeling assumptions at one level (often called the microlevel) to a higher level (the macrolevel). They enable the analysis of macrolevel outcomes and help to transfer models from one scale to another. In general, this could link all levels introduced in Sect. 2.

In this section, we describe different approaches that are used to make this connection. Analytical approaches generally represent groups of individual agents through some macrolevel or average characteristic, often using simplifying assumptions regarding the range of individual agents' characteristics. Simulation approaches describe individual behavior and interactions and then compute the resulting aggregate macroscopic dynamics.

The question of how to aggregate micro-processes to macro-phenomena is not specific to modeling human decision making and behavior. The aggregation of individual behavior and the resulting description of collective action, such as collective motion, is also an ongoing challenge in the natural sciences (Couzin, 2009). Specific assumptions about individual behavior and agent interactions have consequences for the degree of complexity of the macrolevel description. For instance, if agent goals and means do not interact, the properties of single agents can often be added up. If, on the contrary, agents influence each other's goals or interact via the environment, complex aggregate dynamics can arise.

The following sections discuss different aggregation techniques, their underlying assumptions, and how these reflect specific aggregation mechanisms. They are summarized in Table 4.

5.1 Aggregation of preferences: social welfare and voting

Rational choice approaches can also be used to model decision making by agents on higher levels from Table 1, for example firms or countries. The "preferences" of such groups of individuals are often represented by using as the optimization target a social welfare function, which aggregates the members' utility functions either additively ("utilitarian" welfare) or in some nonlinear way to represent inequality aversion (e.g., the Gini–Sen, Atkinson–Theil–Foster, or egalitarian welfare functions; Dagum, 1990). To do so, a common scale of utility must be assumed. For example, individual utility in many economic models equals the logarithm

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

Table 3. Summary table for agent interactions.

Approaches and frameworks	Key considerations	Strengths	Limitations
Classical game theory: strategic interactions between rational agents	What is the game structure (options, possible outcomes, timing, information flow) and what are the players' preferences?	Elegant solutions for low-complexity problems	Difficult to solve for complex games, agents cannot change the rules of the game
Evolutionary game theory: competition and selection between hardwired strategies	What competition and selection mechanisms are there?	Can explain how dominant strategies come about	Agent strategies are modeled as hardwired (no conscious strategy change)
Social influence: agents influence each other's beliefs, preferences, or behaviors	How do influence mechanisms change agent attributes? Is the influence multilateral, dyadic, or directed? How large are deviations?	Allows for the modeling of social learning, preference formation, and herding behavior	Local dynamics are often stylized
Network theory: changing social interaction structures	Is the social network static or adaptive? How much randomness and hierarchy is in the structure? How do agents form new links?	Mathematical formalization to model coevolution of social structure with agent attributes	Micro-interactions mostly dyadic and schematic

of the total monetary value of the individual's consumption. Social welfare functions are indeed used to find optimal policy, for example in cost–benefit analysis (Feldman and Serrano, 2006). Consider a village of farmers growing crops that need different amounts of water so that water management policies affect farmer incomes. The effects of a water policy could then be evaluated using the average, minimal, or average logarithmic income of farmers as a measure of social welfare. The policy option maximizing the chosen indicator should be implemented.

However, it is highly debated whether the utilities of different individuals can really be compared and substituted in the sense that a drop in collective welfare resulting from an actor's decrease in utility can be compensated for by increasing the utility of another actor. Defining suitable group preferences is especially hard when group composition or size changes over time as in intergenerational models (Millner, 2013). Also, in complex organizations, real decisions might be nonoptimal for the group and more explicit models of actual decision procedures may be needed. Models in subfields of game theory (bargaining, voting, or social choice theory) explore the outcomes of formal protocols that are designed to aggregate the group member's heterogeneous preferences. Under different voting or bargaining protocols, subgroups may dominate the decision or the group may be able to reach a compromise (Heitzig and Simmons, 2012). In the above example, the farmers may not agree on a social welfare measure that a policy should optimize but instead on a formal protocol that would allow them to determine a policy for water usage that is acceptable for all.

5.2 Aggregation via markets: economic models and representative agents

A major part of the relevant interaction of contemporary societies with the Earth system is related to the organization of production and consumption on markets. Markets not only mediate between the spheres of production and consumption, but they can also be seen as a mechanism to aggregate agents' decisions and behavior. Economic theory explores how goods and services are allocated and distributed among the various activities (sectors of production) and agents (firms, households, governments) in an economy. Goods and services may be consumed or can be the input factors to economic production. Input factors for production are usually labor and physical capital but can also include financial capital, land, energy, natural resources, and intermediate goods. In markets, the coordination between the demand and supply of goods is mediated through prices that are assumed to reflect information about the scarcity and production costs of goods. Economics compares different kinds of market settings (e.g., auctions, stock exchanges, international trade) with respect to different criteria such as allocative efficiency.

Building on rational choice theory for modeling the decisions of individual agents, microeconomic models in the tradition of neoclassical economics analyze the conditions for an equilibrium between supply and demand on single markets (partial equilibrium theory) and between all markets (general equilibrium theory). The behavior of households and firms is usually modeled as utility maximization

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

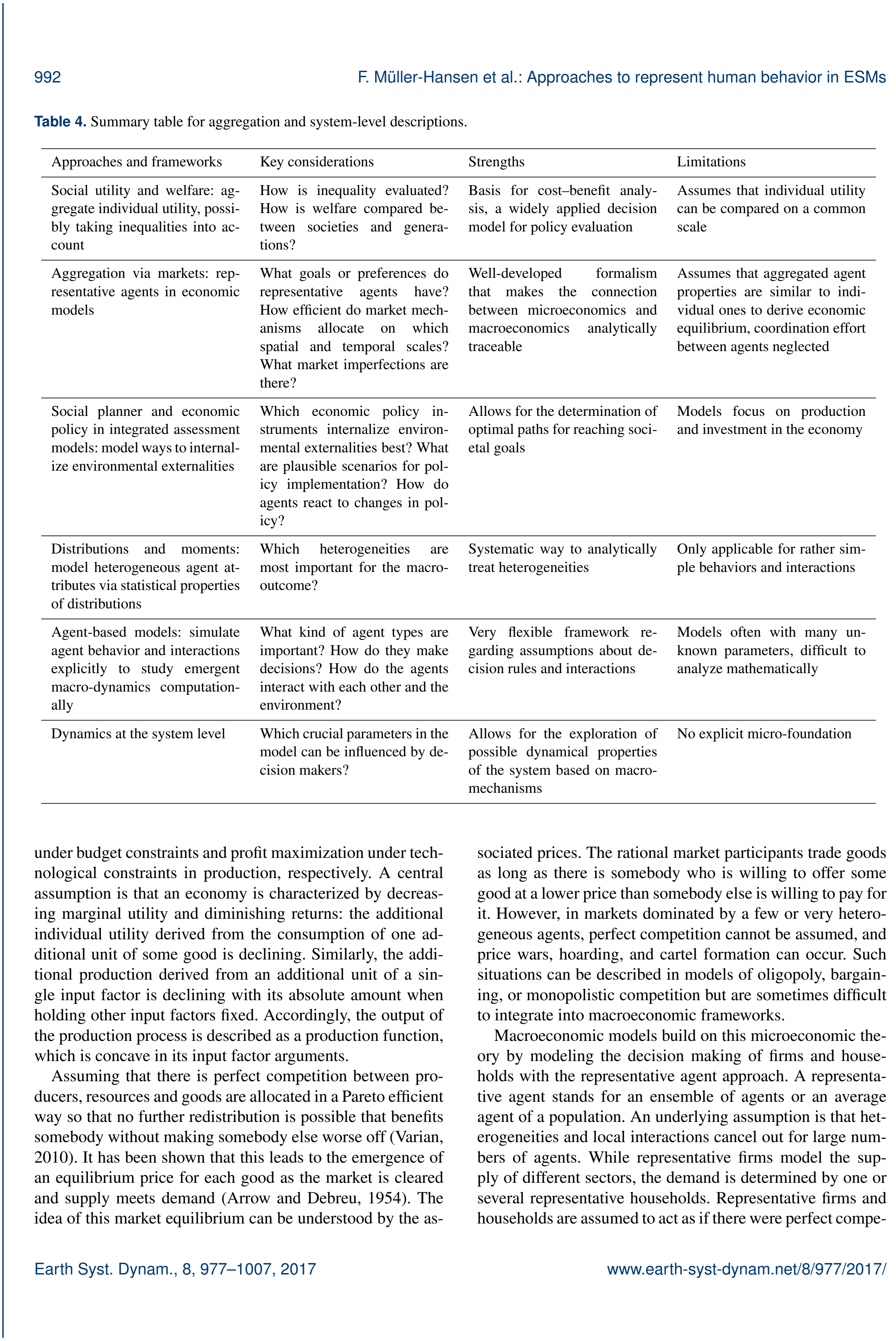


Table 4. Summary table for aggregation and system-level descriptions.

Approaches and frameworks	Key considerations	Strengths	Limitations
Social utility and welfare: aggregate individual utility, possibly taking inequalities into account	How is inequality evaluated? How is welfare compared between societies and generations?	Basis for cost–benefit analysis, a widely applied decision model for policy evaluation	Assumes that individual utility can be compared on a common scale
Aggregation via markets: representative agents in economic models	What goals or preferences do representative agents have? How efficient do market mechanisms allocate on which spatial and temporal scales? What market imperfections are there?	Well-developed formalism that makes the connection between microeconomics and macroeconomics analytically traceable	Assumes that aggregated agent properties are similar to individual ones to derive economic equilibrium, coordination effort between agents neglected
Social planner and economic policy in integrated assessment models: model ways to internalize environmental externalities	Which economic policy instruments internalize environmental externalities best? What are plausible scenarios for policy implementation? How do agents react to changes in policy?	Allows for the determination of optimal paths for reaching societal goals	Models focus on production and investment in the economy
Distributions and moments: model heterogeneous agent attributes via statistical properties of distributions	Which heterogeneities are most important for the macro-outcome?	Systematic way to analytically treat heterogeneities	Only applicable for rather simple behaviors and interactions
Agent-based models: simulate agent behavior and interactions explicitly to study emergent macro-dynamics computationally	What kind of agent types are important? How do they make decisions? How do the agents interact with each other and the environment?	Very flexible framework regarding assumptions about decision rules and interactions	Models often with many unknown parameters, difficult to analyze mathematically
Dynamics at the system level	Which crucial parameters in the model can be influenced by decision makers?	Allows for the exploration of possible dynamical properties of the system based on macro-mechanisms	No explicit micro-foundation

under budget constraints and profit maximization under technological constraints in production, respectively. A central assumption is that an economy is characterized by decreasing marginal utility and diminishing returns: the additional individual utility derived from the consumption of one additional unit of some good is declining. Similarly, the additional production derived from an additional unit of a single input factor is declining with its absolute amount when holding other input factors fixed. Accordingly, the output of the production process is described as a production function, which is concave in its input factor arguments.

Assuming that there is perfect competition between producers, resources and goods are allocated in a Pareto efficient way so that no further redistribution is possible that benefits somebody without making somebody else worse off (Varian, 2010). It has been shown that this leads to the emergence of an equilibrium price for each good as the market is cleared and supply meets demand (Arrow and Debreu, 1954). The idea of this market equilibrium can be understood by the associated prices. The rational market participants trade goods as long as there is somebody who is willing to offer some good at a lower price than somebody else is willing to pay for it. However, in markets dominated by a few or very heterogeneous agents, perfect competition cannot be assumed, and price wars, hoarding, and cartel formation can occur. Such situations can be described in models of oligopoly, bargaining, or monopolistic competition but are sometimes difficult to integrate into macroeconomic frameworks.

Macroeconomic models build on this microeconomic theory by modeling the decision making of firms and households with the representative agent approach. A representative agent stands for an ensemble of agents or an average agent of a population. An underlying assumption is that heterogeneities and local interactions cancel out for large numbers of agents. While representative firms model the supply of different sectors, the demand is determined by one or several representative households. Representative firms and households are assumed to act as if there were perfect compe-

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

tition and they had no market power, i.e., that they optimize their production or consumption taking the prices of goods and production factors as given. The prices of production factors are assumed to equal the value of what they are able to produce additionally by using one additional unit, i.e., their marginal product. In simple macroeconomic models, representative agents interact on perfect markets for all production factors and goods. The solution of the associated optimization problem (with constraints given by a system of nonlinear algebraic equations) specifies the quantity and allocation of input factors, their prices (wages and interest rates), and the production and allocation of consumer goods. A change in one constraint can therefore lead to adjustments in all sectors and new equilibrium prices. For example, in an economy with only two sectors, industry and agriculture, modeled by two representative firms and a representative household, increases in agricultural productivity may lead to the reallocation of labor into the industrial sector and changes in wages.

In reality, prices can undergo rapid fluctuations, which challenges the validity of equilibrium assumptions at least in the short run. Furthermore, production factors may not be fully employed as general equilibrium considerations suggest. Other deviations from efficient equilibria are discussed as market imperfections such as transaction costs, asymmetries in available information, and noncompetitive market structures. Dynamic stochastic general equilibrium (DSGE) models account for the consumption and investment decisions of economic agents under uncertainty and explore the consequences of stochastic shocks on public information or technology for macroeconomic indicators. Many modern DSGE models also incorporate short-term market frictions such as barriers to nominal price adjustments (“sticky” prices) or other market imperfections (Wickens, 2008). However, these models still build on the key concept of general equilibrium because they assume that the state of the economy is always near such an equilibrium and market clearance is fast.

Economic growth models are used to study the long-term dynamics of production and consumption and are therefore an important approach for Earth system modeling. In simple growth models, a homogeneous product is produced per time according to an aggregate production function. A part of the output can be saved as new capital, while the remaining output is consumed. The evolution of the capital stock is given by a differential equation taking into account investments and capital depreciation. In the standard neoclassical growth model, the savings are endogenously determined by the inter-temporal optimization of a representative household and equal investments. The household maximizes an exponentially discounted utility stream (compare Sect. 3.1), which is a function of consumption (Acemoglu, 2009). The central decision of the representative household is how much of the produced output it saves to increase production in the future and therefore cannot consume and enjoy directly. Such inter-temporal optimization problems can be solved either computationally by discretization in time or analytically by applying techniques from optimal control theory[6]. Besides population growth, the only long-term drivers of growth in the standard neoclassical model are exogenously modeled increases in productivity through technological change. In contrast, so-called endogenous growth models exhibit long-run growth and endogenously account for increases in productivity, for example through innovation, human capital, or knowledge accumulation (Romer, 1986; Aghion and Howitt, 1998).

The use of representative agents in macroeconomic models has implications that stem from the implicit assumption that the representative agent has the same properties as an individual of the underlying group (Kirman, 1992; Rizvi, 1994). First, the approach neglects the fact that single agents in the represented group have to coordinate themselves, leaving out problems that arise due to incomplete and asymmetric information. Second, a group of individual maximizers does not necessarily imply collective maximization, challenging the equivalence of the equilibrium outcome. Finally, the representative agent approach may neglect emergent phenomena from heterogeneous micro-interactions (Kirman, 2011).

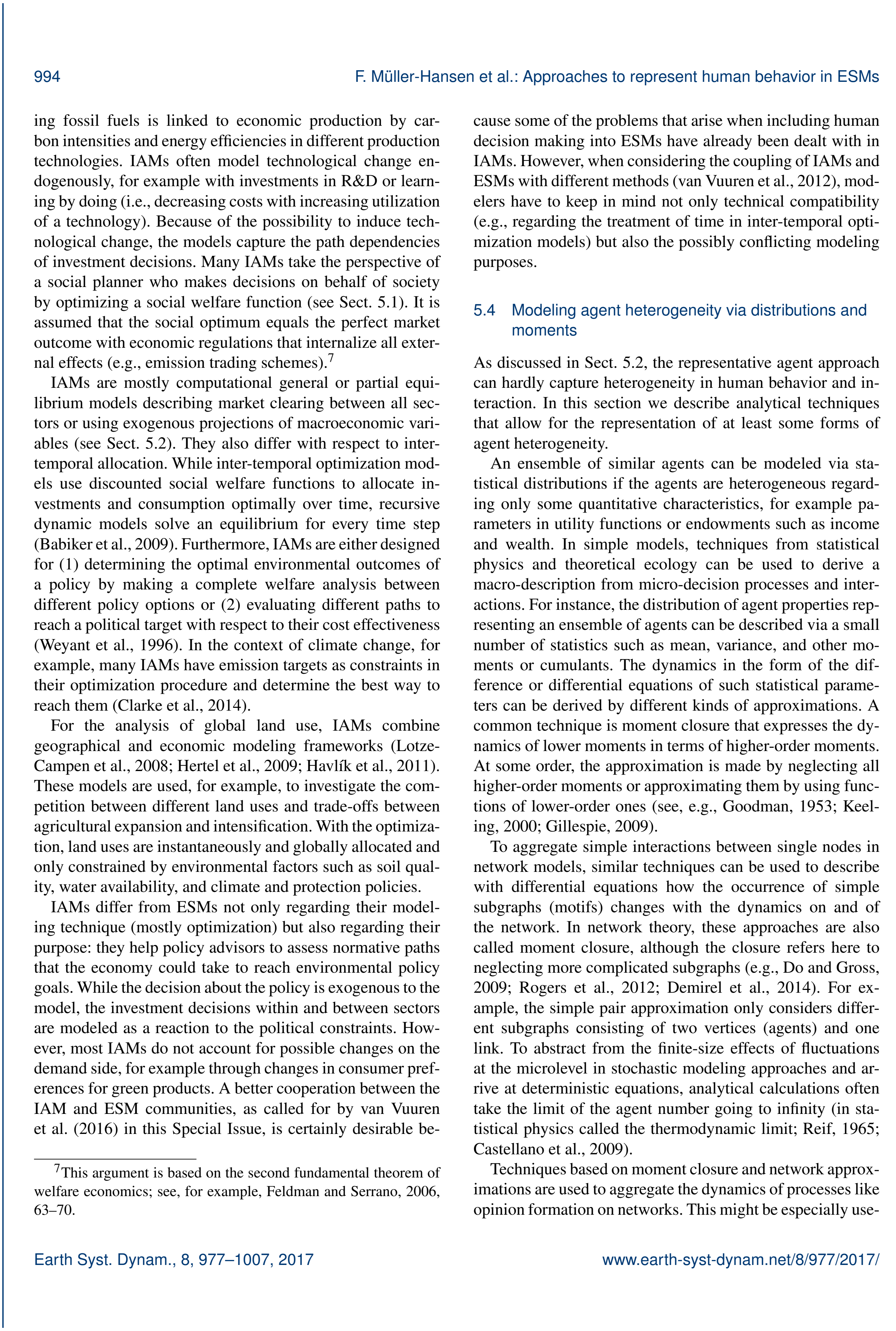
In spite of the deficiencies of the representative agent approach, its application to markets allows for the aggregation of behavior in simple and analytically tractable forms. Modelers who wish to describe economic dynamics at an aggregate level can rely on a well-developed theory that describes many economic phenomena in a good approximation. In the following section, we will discuss how this approach is used to analyze the impacts of economic activities on the environment.

5.3 Modeling of decisions in integrated assessment models: social planner and economic policy

Integrated assessment models (IAMs) comprise a large modeling family that combine economic with environmental dynamics. However, the majority of currently used IAMs draws on ideas from environmental economics. Using the concept of environmental externality, they evaluate the extraction of exhaustible resources, environmental pollution, and overexploitation of ecosystems economically. Externalities are benefits from or damages to the environment that are not reflected in prices and affect other agents in the economy (see, e.g., Perman et al., 2003). These models therefore help to assess economic policies that tackle environmental problems.

State-of-the-art global IAMs combine macroeconomic representations of sectors like the energy and land systems with models of the biophysical bases and environmental impacts of these sectors. For example, CO_2 emitted from burn-

[6] Optimal control theory deals with finding an optimal choice for some control variables (often called policy) of a dynamical system that optimizes a certain objective function using, for example, variational calculus (Kamien and Schwartz, 2012).



ing fossil fuels is linked to economic production by carbon intensities and energy efficiencies in different production technologies. IAMs often model technological change endogenously, for example with investments in R&D or learning by doing (i.e., decreasing costs with increasing utilization of a technology). Because of the possibility to induce technological change, the models capture the path dependencies of investment decisions. Many IAMs take the perspective of a social planner who makes decisions on behalf of society by optimizing a social welfare function (see Sect. 5.1). It is assumed that the social optimum equals the perfect market outcome with economic regulations that internalize all external effects (e.g., emission trading schemes).[7]

IAMs are mostly computational general or partial equilibrium models describing market clearing between all sectors or using exogenous projections of macroeconomic variables (see Sect. 5.2). They also differ with respect to inter-temporal allocation. While inter-temporal optimization models use discounted social welfare functions to allocate investments and consumption optimally over time, recursive dynamic models solve an equilibrium for every time step (Babiker et al., 2009). Furthermore, IAMs are either designed for (1) determining the optimal environmental outcomes of a policy by making a complete welfare analysis between different policy options or (2) evaluating different paths to reach a political target with respect to their cost effectiveness (Weyant et al., 1996). In the context of climate change, for example, many IAMs have emission targets as constraints in their optimization procedure and determine the best way to reach them (Clarke et al., 2014).

For the analysis of global land use, IAMs combine geographical and economic modeling frameworks (Lotze-Campen et al., 2008; Hertel et al., 2009; Havlík et al., 2011). These models are used, for example, to investigate the competition between different land uses and trade-offs between agricultural expansion and intensification. With the optimization, land uses are instantaneously and globally allocated and only constrained by environmental factors such as soil quality, water availability, and climate and protection policies.

IAMs differ from ESMs not only regarding their modeling technique (mostly optimization) but also regarding their purpose: they help policy advisors to assess normative paths that the economy could take to reach environmental policy goals. While the decision about the policy is exogenous to the model, the investment decisions within and between sectors are modeled as a reaction to the political constraints. However, most IAMs do not account for possible changes on the demand side, for example through changes in consumer preferences for green products. A better cooperation between the IAM and ESM communities, as called for by van Vuuren et al. (2016) in this Special Issue, is certainly desirable because some of the problems that arise when including human decision making into ESMs have already been dealt with in IAMs. However, when considering the coupling of IAMs and ESMs with different methods (van Vuuren et al., 2012), modelers have to keep in mind not only technical compatibility (e.g., regarding the treatment of time in inter-temporal optimization models) but also the possibly conflicting modeling purposes.

5.4 Modeling agent heterogeneity via distributions and moments

As discussed in Sect. 5.2, the representative agent approach can hardly capture heterogeneity in human behavior and interaction. In this section we describe analytical techniques that allow for the representation of at least some forms of agent heterogeneity.

An ensemble of similar agents can be modeled via statistical distributions if the agents are heterogeneous regarding only some quantitative characteristics, for example parameters in utility functions or endowments such as income and wealth. In simple models, techniques from statistical physics and theoretical ecology can be used to derive a macro-description from micro-decision processes and interactions. For instance, the distribution of agent properties representing an ensemble of agents can be described via a small number of statistics such as mean, variance, and other moments or cumulants. The dynamics in the form of the difference or differential equations of such statistical parameters can be derived by different kinds of approximations. A common technique is moment closure that expresses the dynamics of lower moments in terms of higher-order moments. At some order, the approximation is made by neglecting all higher-order moments or approximating them by using functions of lower-order ones (see, e.g., Goodman, 1953; Keeling, 2000; Gillespie, 2009).

To aggregate simple interactions between single nodes in network models, similar techniques can be used to describe with differential equations how the occurrence of simple subgraphs (motifs) changes with the dynamics on and of the network. In network theory, these approaches are also called moment closure, although the closure refers here to neglecting more complicated subgraphs (e.g., Do and Gross, 2009; Rogers et al., 2012; Demirel et al., 2014). For example, the simple pair approximation only considers different subgraphs consisting of two vertices (agents) and one link. To abstract from the finite-size effects of fluctuations at the microlevel in stochastic modeling approaches and arrive at deterministic equations, analytical calculations often take the limit of the agent number going to infinity (in statistical physics called the thermodynamic limit; Reif, 1965; Castellano et al., 2009).

Techniques based on moment closure and network approximations are used to aggregate the dynamics of processes like opinion formation on networks. This might be especially use-

[7] This argument is based on the second fundamental theorem of welfare economics; see, for example, Feldman and Serrano, 2006, 63–70.

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

ful in reducing computational complexity when modeling social processes at intermediate levels of aggregation and could allow for the investigation of the interplay of mesoscale social processes with the natural dynamics of the Earth system.

5.5 Aggregation in agent-based models

Agent-based modeling is a computational approach to modeling the emergence of macrolevel or system-level outcomes from microlevel interactions between individual, autonomous agents and between agents and their social and/or biophysical environments (Epstein, 1999; Gilbert, 2008; Edmonds and Meyer, 2013). In agent-based models (ABMs), human behavior is not aggregated to the system level a priori, nor is it assumed that individual behavioral diversity can be represented by a single representative agent as in many macroeconomic models (see Sect. 5.2). Instead, the behavior of heterogeneous agents or groups of agents is explicitly simulated to study the resulting aggregate outcomes. As each action of an individual agent is interdependent, i.e., it depends on the decisions or actions of other agents within structures such as networks or space, local interactions can give rise to complex, emergent patterns of aggregate behavior at the macrolevel (Page, 2015). ABMs allow for the exploration of such nonlinear behavior in order to understand possible future developments of the system or assess possible unexpected outcomes of disturbances or policy interventions. Agent-based modeling is widely used to study complex systems in computational social science (Conte and Paolucci, 2014), land-use science (Matthews et al., 2007), political science (de Marchi and Page, 2014), computational economics (Tesfatsion, 2006; Heckbert et al., 2010; Hamill and Gilbert, 2016), social–ecological systems research (Schlüter et al., 2012; An, 2012), and ecology (Grimm and Railsback, 2005), among others.[8]

Agents in ABMs can be individuals, households, firms, or other collective actors, as well as other entities or groups thereof, such as fish, fish populations, or plant functional types. Agents are assumed to be diverse and heterogeneous; i.e., they can belong to different types and can vary within one type, respectively. Agent types can be characterized by different attributes and decision-making models (e.g., large and commercial versus small and traditional farms). Heterogeneity within a type is often represented through quantitative differences in the values of these attributes (e.g., regarding market access, social, or financial capital). The decision making and behavior of the agents can be modeled with any of the approaches introduced in Sect. 3 or can be based on data or observations that are formalized in equations, decision trees, or other formal rules. In empirical ABMs, agents are often classified into empirically based agent types, which are characterized by attributes and decision heuristics derived from empirical data obtained through interviews or surveys (Smajgl and Barreteau, 2014). Increasingly, social science theories of human behavior beyond the rational actor are being used in ABMs to represent more realistic human decision making. However, many challenges remain to translate these theories for usage in ABMs (Schlüter et al., 2017).

Probabilistic and stochastic processes are often used to capture uncertainty in and the impact of random events on human decision making and assess the consequences for macrolevel outcomes. For example, random events at the local level, such as a random encounter between two agents that results in a strategy change of one agent or a system-level environmental variation, can give rise to nonlinear macrodynamics such as a sudden shift into a different system state (Schlüter et al., 2016).

In addition to the behavior of the agents, ABMs of human–environment systems incorporate the dynamics of the biophysical environment resulting from natural processes and human actions insofar as it is relevant for the agents' behavior and to understand feedbacks between human behavior and environmental processes. For example, in an ABM by Martin et al. (2016), a number of cattle ranchers can move their livestock between grassland patches in a landscape. Overgrazing in one year decreases feed availability in the following year because of the underlying biomass regrowth dynamics. Agents decide how many cattle to graze on a particular land patch based on their individual goals or needs, information on the state of the grassland, beliefs about the future, and interactions with other ranchers. The model can reveal the interplay and success of different land-use strategies on common land and assess their vulnerability to shocks such as droughts. Most ABMs in the context of land-use science have so far been developed for local or regional study areas, taking into account local specificities and fitting behavioral patterns to data acquired in the field (Parker et al., 2003; Matthews et al., 2007; Groeneveld et al., 2017). They are often combined with cellular automaton models that describe the dynamics and state of the physical land system (e.g., Heckbert, 2013). In these ABMs, the spatial embedding of agents usually plays an important role (Stanilov, 2012).

Because ABMs can integrate a diversity of individual decision making, heterogeneity of actors, and interactions between agents constrained by social networks or space and social and environmental processes, they are particularly suitable to study feedbacks between human action and biophysical processes. In the context of ESM these may include human adaptive responses to environmental change, such as the effects of climate change on agriculture and water availability, to policies such as bioenergy production or the global consequences of shifts in diets in particular regions. Agent-based modeling is also a useful tool to unravel the causal mechanisms underlying system-level phenomena (Epstein, 1999; Hedström and Ylikoski, 2010) and thus enhance the

[8] Note that in some scientific communities, this class of modeling approaches is also known as multi-agent simulation (MAS; Bousquet and Le Page, 2004) or individual-based modeling (Grimm and Railsback, 2005).

understanding of key human–environment interactions that may give rise to observed Earth system dynamics. However, because of their potentially high complexity and dimensionality in state and parameter space, ABMs are often difficult to analyze and may require high computational capacities and sophisticated model analysis techniques to understand their dynamics beyond single trajectories.

Agent-based approaches can be applied without modeling each individual agent explicitly. It suffices to model a representative statistical sample of agents that depicts the important heterogeneities of the underlying population. To capture major types of human behavior, a recent proposal involves agent functional types based on a theoretically derived typology of agent attributes, interactions, and roles (Arneth et al., 2014). This proposal is explored for modeling the adaptation of land-use practices to climate change impacts (Murray-Rust et al., 2014). Agent functional types represent a typology that is theoretically constructed instead of data driven, which is common in empirically based ABMs. Agent-based approaches are promising for Earth system modeling because they allow modelers to address questions of interactions across levels, for instance how global patterns of land use emerge from interdependent regional and local land-use decisions, which are in turn constrained by the emerging global patterns. Furthermore, they would allow for the integration of uncertainty, agent heterogeneity, and the aggregation of detailed technological and environmental changes (Farmer et al., 2015).

5.6 Dynamics at the system level: system dynamics, stock-flow consistent, and input–output models

This final subsection discusses modeling approaches without explicit micro-foundations. Decisions in such models are not modeled explicitly with one of the options discussed in Sect. 3 but, as policy decisions in integrated assessment models, through the construction of different scenarios for the evolution of crucial exogenous parameters in the model.

Global system dynamics models describe the economy, population, and crucial parts of the Earth system and their dynamic interactions at the level of aggregate dynamic variables, usually modeling the dynamics as ordinary differential equations or difference equations to project future developments. The equations are often built on stylized facts about the dynamics of the underlying subsystems and are linked by functions with typically many parameters. Modelers employ system dynamics models to develop scenarios based on different sets of model parameters and assess the system stability and transient dynamics. In comparison to equilibrium approaches, system dynamics models capture the inertia of socioeconomic systems at the cost of a higher dimensional parameter space. This can lead to more complex dynamics like oscillations or overshooting. System dynamics models can be very detailed, like the World3 model commissioned by the Club of Rome for their famous report "Limits to Growth" (Meadows et al., 1972, 2004), the GUMBO model (Boumans et al., 2002), or the International Futures model (Hughes, 1999). Subsystems of such models comprise the human population (sometimes disaggregated between regions and age groups), the agricultural and industrial sector, and the state of the environment (pollution and resource availability). Simpler models describe the dynamics of only a few aggregated variables at the global level (Kellie-Smith and Cox, 2011) or confined to a region (Brander and Taylor, 1998).

Other system-level approaches to macroeconomic modeling emphasize self-reinforcing processes in the economy and point at positive feedback mechanisms, resulting in multistability or even instability (e.g., increasing returns to scale in production and self-amplification of expectations during economic bubbles). For example, post-Keynesian economists use stock-flow consistent models to track the complete monetary flows in an economy in which low aggregate demand can lead to underutilization of production factors and the state plays an active role to stabilize the economy. In these models, a social accounting matrix provides a detailed framework of transactions (e.g., monetary flows) between households, firms, and the government, which hold stocks of assets and commodities (Godley and Lavoie, 2007).

Input–output models track flows to much more detail between different industries or sectors of production (Leontief, 1986; Ten Raa, 2005; Miller and Blair, 2009). Each industry or production process is modeled by a "Leontief" production function, which is characterized by fixed proportions of input factors that depend on the available technology. For example, an input–output model can describe which input factors, such as land, fertilizer, machinery, irrigation water, and labor, are required for satisfying the demand of an agricultural commodity with a mix of production techniques. The model would consider that some of these inputs have to be produced themselves using other types of inputs. Outputs also include unwanted side products, such as manure in cattle production. Such models are used, for instance, to explore how changes in demand would lead to higher-order effects along the supply chain. Regional input–output models also account for spatial heterogeneity and are used, for example, to evaluate the possible impacts of extreme climate events on the global supply chain (Bierkandt et al., 2014).

While the approaches discussed above focus on the monetary dimension of capital and goods, models from ecological economics (van den Bergh, 2001) track material flows or integrate material with financial accounting. For example, input–output modeling has been extended to analyze industrial metabolism, i.e., material and energy flows and their environmental impacts in modern economies (Fischer-Kowalski and Haberl, 1997; Ayres and Ayres, 2002; Suh, 2009). Regionalized versions of such models can, for instance, be used to estimate the environmental footprint that industrialized countries have in other regions (Wiedmann, 2009). In the emerging field of ecological macroeconomics (see Hardt and O'Neill, 2017, for a detailed review of mod-

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

eling approaches), stock-flow consistent and input–output models have been combined into one framework for tracking financial and material flows (Berg et al., 2015). Other ecological models use the flow–fund approach by Georgescu-Roegen (1971) or combine it with stock-flow consistent modeling approaches (Dafermos et al., 2017). While the flow concept refers to a stock per time, a fund is the potentiality of a system to provide a service. The important difference lies in the observation that a stock can be depleted or accumulated in one time step, while a fund can provide its service only once per time step. This distinction reflects physical constraints on the production process that have important consequences for modeling the social metabolism. Garrett (2015) and Jarvis et al. (2015) in this Special Issue provide an extreme view on the dynamics of social metabolism based only on thermodynamic considerations without taking human decision making or agency into account.

In order to make approaches that only consider the system level useful for modeling the impact of humans on the Earth system, they could be combined with approaches that model the development of new production technologies and how the deployment of new technologies is affected by decisions at different levels (consumers, firms, and governments). Even if this integration with decision models proves difficult, the approaches discussed in this section can help link social and environmental dynamics in new ways, providing an important methodology to include humans into ESMs.

6 Discussion

In the previous three sections, we showed that there is a diversity of approaches to model individual human decision making and behavior, to describe interactions between agents, and to aggregate these processes. The discussion of strengths and limitations of the modeling approaches showed possible underlying assumptions and connections to theories of human behavior. While some modeling techniques are compatible with many theories of human behavior or decision making and can thus be used with a variety of assumptions, other techniques significantly constrain possible assumptions.

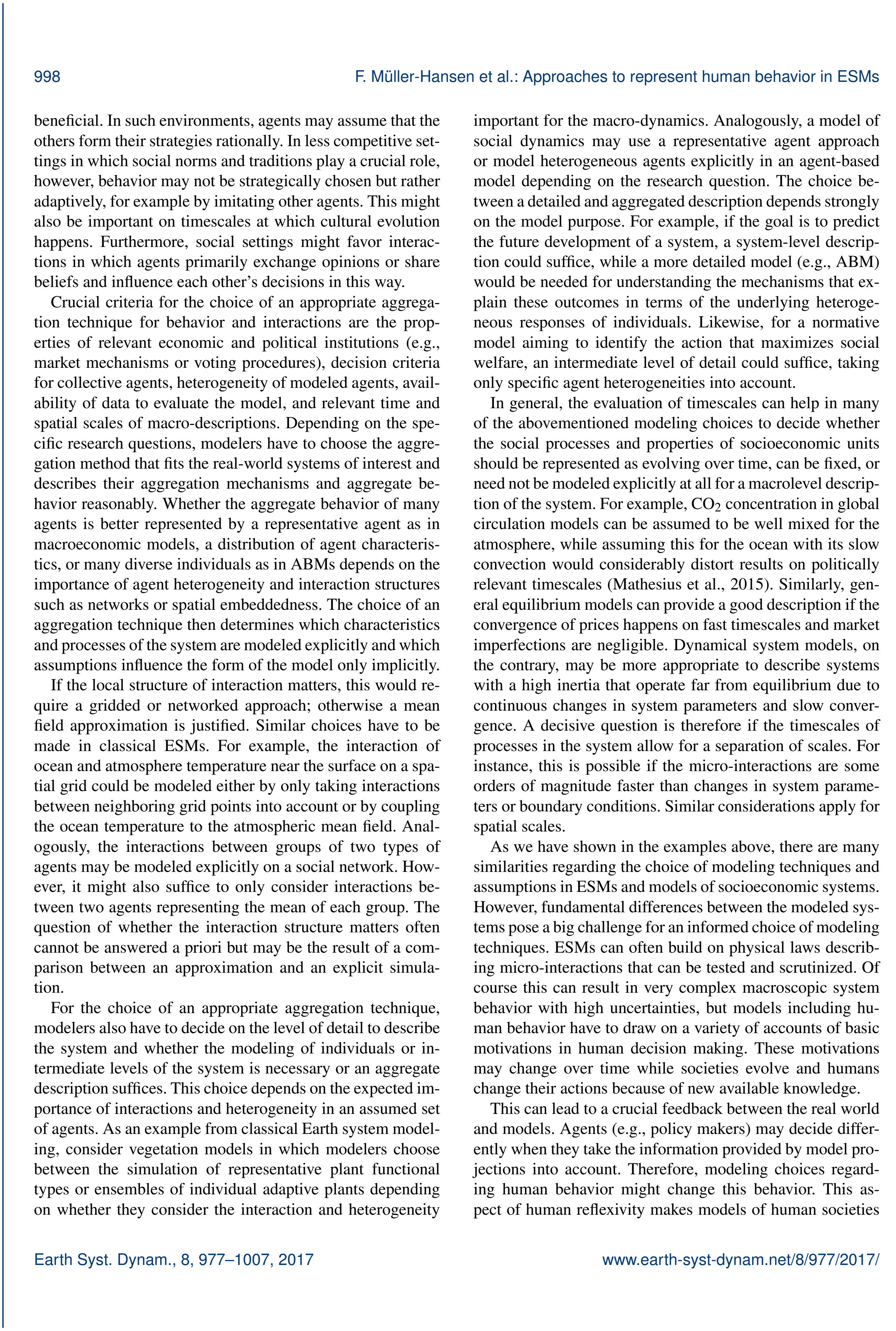
For many relevant questions in global environmental change research, a dynamical representation of humans in ESMs may not be necessary. If behavioral patterns are not expected to change over the relevant timescales or feedbacks between natural and social dynamics are sufficiently weak, modelers can simply use conventional scenario approaches.

However, if behavioral patterns are expected to change over time and give rise to strong feedbacks with the environment, then an explicit representation of human decision making will provide new insights into the joint dynamics. In this case, modelers have to carefully choose which assumptions about human behavior and decision making are plausible for their specific modeling purpose. Modeling choices require a constant interplay between model development and the research questions that drive it.

Because there is no general theory of human decision making and behavior, especially not for social collectives, we cannot provide a specific recipe for including humans into ESMs. In Table 5, we summarize the approaches we discussed in this paper and collect important questions to guide the choice of appropriate model assumptions and approaches. To find the right assumptions for a specific context, modelers can further build on and consult existing social scientific research, even though ambiguities due to a fragmentation of the literature between opposing schools of thought and difficulties in generalizing single case studies from their local or cultural specificities can make some of the research difficult to access. In case of doubt, modelers can team up with social scientists to conduct empirical research in the specific context needed to select the appropriate approach. The selection of a modeling technique compatible with the chosen assumptions also has to consider its limitations for meaningfully answerable research questions and the analyses that it can provide. In the following, we discuss some important considerations regarding individual decision making, interactions, and aggregation.

Concerning individual agents, we identified three important determinants in decision models: motives, restrictions, and decision rules. Modelers need to take the many factors into account that influence which assumptions about each of these three determinants are applicable in a given context. For instance, modelers can make different assumptions about whether agents only consider financial incentives or also take into account other criteria, such as a desire for fair outcome distributions (Opp, 1999), depending on whether a situation is more or less competitive or cooperative. Research shows that the relevance of motives and goals can vary over time and that surprisingly subtle cues can change their importance (Lindenberg, 1990; Tversky and Kahneman, 1985). Likewise, the choice of a plausible decision rule depends on the studied context. For instance, a decision rule that requires complex computations may be relatively plausible in contexts in which agents make decisions with important consequences and in which they have the information and time needed to compare alternatives. When stakes are low and time to decide is limited, however, more simple decision rules are certainly more plausible. Cognitively demanding decision rules are also more plausible when decision makers are collectives, such as companies and governments. Sometimes, it may even be reasonable to assume that agents use combinations of different decision models (Camerer and Ho, 1999).

Important criteria for choosing an appropriate model of agent interactions are the type and setting of interactions, the assumptions that agents make about each other, the influence they may exert on each other, and the structure of interactions. For example, interactions in competitive environments will only lead to cooperation if this is individually



beneficial. In such environments, agents may assume that the others form their strategies rationally. In less competitive settings in which social norms and traditions play a crucial role, however, behavior may not be strategically chosen but rather adaptively, for example by imitating other agents. This might also be important on timescales at which cultural evolution happens. Furthermore, social settings might favor interactions in which agents primarily exchange opinions or share beliefs and influence each other's decisions in this way.

Crucial criteria for the choice of an appropriate aggregation technique for behavior and interactions are the properties of relevant economic and political institutions (e.g., market mechanisms or voting procedures), decision criteria for collective agents, heterogeneity of modeled agents, availability of data to evaluate the model, and relevant time and spatial scales of macro-descriptions. Depending on the specific research questions, modelers have to choose the aggregation method that fits the real-world systems of interest and describes their aggregation mechanisms and aggregate behavior reasonably. Whether the aggregate behavior of many agents is better represented by a representative agent as in macroeconomic models, a distribution of agent characteristics, or many diverse individuals as in ABMs depends on the importance of agent heterogeneity and interaction structures such as networks or spatial embeddedness. The choice of an aggregation technique then determines which characteristics and processes of the system are modeled explicitly and which assumptions influence the form of the model only implicitly.

If the local structure of interaction matters, this would require a gridded or networked approach; otherwise a mean field approximation is justified. Similar choices have to be made in classical ESMs. For example, the interaction of ocean and atmosphere temperature near the surface on a spatial grid could be modeled either by only taking interactions between neighboring grid points into account or by coupling the ocean temperature to the atmospheric mean field. Analogously, the interactions between groups of two types of agents may be modeled explicitly on a social network. However, it might also suffice to only consider interactions between two agents representing the mean of each group. The question of whether the interaction structure matters often cannot be answered a priori but may be the result of a comparison between an approximation and an explicit simulation.

For the choice of an appropriate aggregation technique, modelers also have to decide on the level of detail to describe the system and whether the modeling of individuals or intermediate levels of the system is necessary or an aggregate description suffices. This choice depends on the expected importance of interactions and heterogeneity in an assumed set of agents. As an example from classical Earth system modeling, consider vegetation models in which modelers choose between the simulation of representative plant functional types or ensembles of individual adaptive plants depending on whether they consider the interaction and heterogeneity important for the macro-dynamics. Analogously, a model of social dynamics may use a representative agent approach or model heterogeneous agents explicitly in an agent-based model depending on the research question. The choice between a detailed and aggregated description depends strongly on the model purpose. For example, if the goal is to predict the future development of a system, a system-level description could suffice, while a more detailed model (e.g., ABM) would be needed for understanding the mechanisms that explain these outcomes in terms of the underlying heterogeneous responses of individuals. Likewise, for a normative model aiming to identify the action that maximizes social welfare, an intermediate level of detail could suffice, taking only specific agent heterogeneities into account.

In general, the evaluation of timescales can help in many of the abovementioned modeling choices to decide whether the social processes and properties of socioeconomic units should be represented as evolving over time, can be fixed, or need not be modeled explicitly at all for a macrolevel description of the system. For example, CO_2 concentration in global circulation models can be assumed to be well mixed for the atmosphere, while assuming this for the ocean with its slow convection would considerably distort results on politically relevant timescales (Mathesius et al., 2015). Similarly, general equilibrium models can provide a good description if the convergence of prices happens on fast timescales and market imperfections are negligible. Dynamical system models, on the contrary, may be more appropriate to describe systems with a high inertia that operate far from equilibrium due to continuous changes in system parameters and slow convergence. A decisive question is therefore if the timescales of processes in the system allow for a separation of scales. For instance, this is possible if the micro-interactions are some orders of magnitude faster than changes in system parameters or boundary conditions. Similar considerations apply for spatial scales.

As we have shown in the examples above, there are many similarities regarding the choice of modeling techniques and assumptions in ESMs and models of socioeconomic systems. However, fundamental differences between the modeled systems pose a big challenge for an informed choice of modeling techniques. ESMs can often build on physical laws describing micro-interactions that can be tested and scrutinized. Of course this can result in very complex macroscopic system behavior with high uncertainties, but models including human behavior have to draw on a variety of accounts of basic motivations in human decision making. These motivations may change over time while societies evolve and humans change their actions because of new available knowledge.

This can lead to a crucial feedback between the real world and models. Agents (e.g., policy makers) may decide differently when they take the information provided by model projections into account. Therefore, modeling choices regarding human behavior might change this behavior. This aspect of human reflexivity makes models of human societies

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

Table 5. Collection of questions that may guide the choice of modeling approaches and assumptions.

Category	Important modeling questions
Modeling individual decision making and behavior	What goals do agents pursue? What constraints do they have? What decision rules do agents use? How do agents acquire information and beliefs about their environment?
Modeling interactions between agents	Do agents interact in a competitive environment, or are interactions primarily governed by social norms? What do agents assume about each other's rationality? Do agents choose actions strategically or adaptively? How are agents influenced by others regarding their beliefs and norms? What structure do the interactions have, and how does the structure evolve?
Aggregating behavior and modeling dynamics at the system level	Are decisions aggregated through political institutions (e.g., voting procedures) or markets? According to what criteria do policy makers decide, and what controls do they have? Is the heterogeneity of agent characteristics and interactions important? Which macrolevel measures are dynamic and which can be assumed to be fixed?

fundamentally different from natural science models and is closely linked to the important difference in social modeling between normative and descriptive model purposes. For example, models that optimize social welfare usually reflect the goal that a government should pursue and therefore have a normative purpose. However, if this model is used to guide policy making while taking into account the actual and perceived controls of policy makers and considers the effect of compromises between different interest groups, it could also describe its behavior. This example shows the often intricate interconnections between normative and descriptive assumptions in decision modeling that modelers should be aware of.

This is further complicated by the observation that the same assumption may be understood in one model as a descriptive (positive) statement, whereas in another model it may be meant as a prescriptive (normative) one. For example, in a model of agricultural markets, the assumption that big commercial farms maximize their profits might be a reasonable descriptive approximation. In contrast, in a model that asks how smallholder farms could survive under competitive market conditions, the same assumption gets a strong normative content.

Another difficulty is that model choices are often not only based on the most plausible assumptions about human decision making but are also strongly influenced by considerations about the assumption's mathematical convenience. Choosing assumptions for technical reasons, for example mathematical simplicity and tractability, may be problematic because it remains unexplained how they are related to the real world. Because not all assumptions can be easily implemented in formal models, a trade-off often has to be found between the plausibility and technical practicality of the assumptions.

Most of the global models reviewed here that describe human interactions with the Earth system are based on economic assumptions about the behavior of humans and societies. They are often only linked in a one-way fashion to the biogeophysical part of the Earth system. Including closed feedback loops between social and environmental dynamics into ESMs is still a big challenge. To advance this endeavor, more work is needed to synthesize modeling approaches that can represent various aspects of human behavior in the context of global modeling, even if the need for generalizations and the formalization of human behavior is sometimes met with skepticism or rejection by social scientists who emphasize the context dependence and idiosyncrasy of human behavior. Of course, models that use simple theories of human decision making and behavior to describe human–environment interactions in the global context cannot claim to capture all real-world social interactions. If models considered the heterogeneity of agents in all relevant aspects, they would have to be much more complex than all models that have been developed to date. However, in many real-life settings, even simple conceptual models of social mechanisms are good descriptions of the key features of the dynamics at work, as we have highlighted throughout this review. Including such formal descriptions of idealized social mechanisms can therefore be a good starting point for understanding feedbacks in the Earth system and their qualitative consequences, which have so far not been considered explicitly in global models.

7 Summary and conclusion

In this review, we discussed common modeling techniques and theories that could be potentially used to include human decision making and the resulting feedbacks with environmental dynamics into Earth system models (ESMs). Although we could only discuss the basic aspects of the presented modeling techniques, it is apparent that modelers who want to include humans into ESMs are confronted with crucial choices of which assumptions to make about human behavior and which appropriate techniques to use.

As Table 5 summarizes, we discussed techniques and modeling assumptions in three different categories. First, individual decision modeling focuses on decision processes and the resulting behavior of single agents and therefore has to make assumptions about the determinants of choices between behavioral options. Second, models of interactions between agents capture how decisions depend upon each other and how agents influence each other regarding different decision criteria. Third, modeling techniques that aggregate agent behavior and interactions to a system-level description are crucial for modeling human behavior at scales relevant for the Earth system and require ingredients from the first and second categories. To include human decision making into ESMs, techniques and assumptions from these three categories have to be combined. Finally, we discussed important questions regarding the choice of modeling approaches and their interrelation with assumptions about human behavior and decision making, for example regarding the level of description and the relevant timescales but also the difficulties that can arise due to human reflexivity and the amalgamation of normative and descriptive assumptions in models.

Most formal models that describe human behavior in global environmental contexts are based on economic approaches. This is not surprising because many human interactions with the environment are driven by economic forces, and economics has a stronger focus on formal models than other social sciences. However, we think that it is necessary to advance research that builds on insights from other social sciences and applies social modeling and simulation in the context of global environmental change. One important aim of such research would be to provide a theoretical basis for including processes of social evolution and institutional development into ESMs. If we want to explore the possible futures of the Earth, we need to get a better understanding of how the long-term dynamics of the Earth system are shaped by these cultural and social processes.

A new generation of ESMs can build on various approaches, some of which we reviewed here, to include human decision making and behavior explicitly into Earth system dynamics. However, ambitious endeavors like this have to take into account that the modeling of human behavior and social processes is a contested topic, and the assumptions and corresponding modeling techniques need to be chosen carefully with an awareness of their strengths and limitations for the specific modeling purpose.

Data availability. No data sets were used in this article.

Competing interests. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue "Social dynamics and planetary boundaries in Earth system modelling". It is not associated with a conference.

Acknowledgements. The idea for this paper emerged during discussions at the LOOPS workshop in 2014. We thank all the participants for their contributions. Furthermore, we thank the referees and the editor for helpful comments that improved the paper. We thank Rainer Hegselmann, Silvana Tiedemann, Tim Kittel, Wolfram Barfuss, Nicola Botta, and Felix Kersting for their comments on previous versions of the paper. Finn Müller-Hansen acknowledges funding by the DFG (IRTG 1740/TRP 2011/50151-0). Maja Schlüter acknowledges funding by the European Research Council under the European Union Seventh Framework Programme (FP/2007-2013)/ERC grant agreement no. 283950, SES-LINK, and a core grant to the Stockholm Resilience Centre by Mistra. Jonathan F. Donges thanks the Stordalen Foundation (via the Planetary Boundary Research Network PB.net), the EarthDoc network, and the Leibniz Association (project DOMINOES) for financial support.

Edited by: Sarah Cornell
Reviewed by: Andrew Jarvis and one anonymous referee

References

Acemoglu, D.: Introduction to Modern Economic Growth, Princeton University Press, Princeton, NJ, 2009.

Ackermann, K. A., Fleiß, J., and Murphy, R. O.: Reciprocity as an individual difference, J. Conflict Resolut., 60, 340–367, https://doi.org/10.1177/0022002714541854, 2016.

Aghion, P. and Howitt, P.: Endogenous Growth Theory, MIT Press, Cambridge, Massachusetts and London, UK, 1998.

Ainslie, G. and Haslam, N.: Hyperbolic Discounting, in: Choice over time, edited by: Loewenstein, G. and Elster, J., Russell Sage Foundation, New York, 57–92, 1992.

Akers, R. L., Krohn, M. D., Lanza-Kaduce, L., and Radosevich, M.: Social Learning and Deviant Behavior: A specific Test of a general Theory, Am. Sociol. Rev., 44, 636–655, https://doi.org/10.2307/2094592, 1979.

An, L.: Modeling human decisions in coupled human and natural systems: Review of agent-based models, Ecol. Model., 229, 25–36, https://doi.org/10.1016/j.ecolmodel.2011.07.010, 2012.

Arneth, A., Brown, C., and Rounsevell, M. D. A.: Global models of human decision-making for land-based mitigation and adaptation assessment, Nature Climate Change, 4, 550–557, https://doi.org/10.1038/nclimate2250, 2014.

Arrow, K. J. and Debreu, G.: Existence of an Equilibrium for a Competitive Economy, Econometrica, 22, 265–290, https://doi.org/10.2307/1907353, 1954.

Auer, S., Heitzig, J., Kornek, U., Schöll, E., Kurths, J., Scholl, E., Kurths, J., Schöll, E., and Kurths, J.: The Dynamics of Coalition Formation on Complex Networks, Nature Scientific Reports, 5, 13386, https://doi.org/10.1038/srep13386, 2015.

Aumann, R. J.: War and peace, P. Natl. Acad. Sci. USA, 103, 17075–17078, https://doi.org/10.1073/pnas.0608329103, 2006.

Axelrod, R.: The evolution of cooperation, Basic Books, New York, 1984.

Axelrod, R.: The dissemination of culture: A model with local convergence and global polarization, J. Conflict Resolut., 41, 203–226, https://doi.org/10.1177/0022002797041002001, 1997.

Ayres, R. U. and Ayres, L. (Eds.): A Handbook of Industrial Ecology, Edward Elgar, Cheltenham, UK, 2002.

Babad, E. and Katz, Y.: Wishful Thinking – Against All Odds, J. Appl. Soc. Psychol., 21, 1921–1938, https://doi.org/10.1111/j.1559-1816.1991.tb00514.x, 1991.

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

Babiker, M., Gurgel, A., Paltsev, S., and Reilly, J.: Forward-looking versus recursive-dynamic modeling in climate policy analysis: A comparison, Econ. Model., 26, 1341–1354, https://doi.org/10.1016/j.econmod.2009.06.009, 2009.

Baker, W. L.: A review of models of landscape change, Landscape Ecol., 2, 111–133, https://doi.org/10.1007/bf00137155, 1989.

Balint, T., Lamperti, F., Mandel, A., Napoletano, M., Roventini, A., and Sapio, A.: Complexity and the Economics of Climate Change: A Survey and a Look Forward, Ecol. Econ., 138, 252–265, https://doi.org/10.1016/j.ecolecon.2017.03.032, 2017.

Balke, T. and Gilbert, N.: How Do Agents Make Decisions? A Survey, JASSS-J. Artif. Soc. S., 17, 13, https://doi.org/10.18564/jasss.2687, 2014.

Barfuss, W., Donges, J. F., Wiedermann, M., and Lucht, W.: Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dynam., 8, 255–264, https://doi.org/10.5194/esd-8-255-2017, 2017.

Bellman, R.: A Markovian decision process, Indiana U. Math. J., 6, 679–684, https://doi.org/10.1512/iumj.1957.6.56038, 1957.

Berg, M., Hartley, B., and Richters, O.: A stock-flow consistent input-output model with applications to energy price shocks, interest rates, and heat emissions, New J. Phys., 17, 15011, https://doi.org/10.1088/1367-2630/17/1/015011, 2015.

Bierkandt, R., Wenz, L., Willner, S. N., and Levermann, A.: Acclimate - a model for economic damage propagation. Part I: basic formulation of damage transfer within a global supply network and damage conserving dynamics, Environment Systems and Decisions, 34, 507–524, https://doi.org/10.1007/s10669-014-9523-4, 2014.

Bikhchandani, S., Hirshleifer, D., and Welch, I.: A Theory of Fads, Fashion, Custom, and Cultural-Change as Informational Cascades, J. Polit. Econ., 100, 992–1026, https://doi.org/10.1086/261849, 1992.

Boccaletti, S., Bianconi, G., Criado, R., del Genio, C. I., Gómez-Gardeñes, J., Romance, M., Sendiña-Nadal, I., Wang, Z., and Zanin, M.: The structure and dynamics of multilayer networks, Phys. Rep., 544, 1–122, https://doi.org/10.1016/j.physrep.2014.07.001, 2014.

Boudon, R.: The Logic of Social Action. An Introduction to Sociological Analysis, Routledge & Kegan Paul, London, 1981.

Boumans, R., Costanza, R., Farley, J., Wilson, M. A., Portela, R., Rotmans, J., Villa, F., and Grasso, M.: Modeling the dynamics of the integrated earth system and the value of global ecosystem services using the GUMBO model, Ecol. Econ., 41, 529–560, https://doi.org/10.1016/S0921-8009(02)00098-8, 2002.

Bousquet, F. and Le Page, C.: Multi-agent simulations and ecosystem management: a review, Ecol. Model., 176, 313–332, https://doi.org/10.1016/j.ecolmodel.2004.01.011, 2004.

Bowles, S.: Endogenous Preferences: The Cultural Consequences of Markets and Other Economic Institutions, J. Econ. Lit., 36, 75–111, 1998.

Brander, J. A. and Taylor, M. S.: The Simple Economics of Easter Island: A Ricardo-Malthus Model of Renewable Resource Use, A. Econ. Rev., 88, 119–138, 1998.

Brown, C., Brown, K., and Rounsevell, M.: A philosophical case for process-based modelling of land use change, Modeling Earth Systems and Environment, 2, 50, https://doi.org/10.1007/s40808-016-0102-1, 2016.

Brown, C., Alexander, P., Holzhauer, S., and Rounsevell, M. D.: Behavioural models of climate change adaptation and mitigation in land-based sectors, WIREs Climate Change, 8, e448, https://doi.org/10.1002/wcc.448, 2017.

Brown, D. G., Walker, R., Manson, S., and Seto, K.: Modeling Land-Use and Land-Cover Change, in: Land Change Science: Observing, Monitoring and Understanding Trajectories of Change on the Earth's Surface, edited by: Gutman, G., Janetos, A. C., Justice, C. O., Moran, E. F., Mustard, J. F., Rindfuss, R. R., Skole, D., Turner II, B. L., and Cochrane, M. A., Springer, Dordrecht, the Netherlands, chap. 23, 395–409, https://doi.org/10.1007/978-1-4020-2562-4_23, 2004.

Bruhin, A., Fehr-Duda, H., and Epper, T.: Risk and Rationality: Uncovering Heterogeneity in Probability Distortion, Econometrica, 78, 1375–1412, https://doi.org/10.3982/ECTA7139, 2010.

Camerer, C. and Ho, T. H.: Experience-weighted attraction learning in normal form games, Econometrica, 67, 827–874, https://doi.org/10.1111/1468-0262.00054, 1999.

Castellano, C., Fortunato, S., and Loreto, V.: Statistical physics of social dynamics, Rev. Mod. Phys., 81, 591–646, https://doi.org/10.1103/RevModPhys.81.591, 2009.

Chong, E. K. P. and Zak, S. H.: An Introduction to Optimization, 4th Edn., Wiley, Hoboken, NJ, 2013.

Clarke, L., Jiang, K., Akimoto, K., Babiker, M., Blanford, G., Fisher-Vanden, K., Hourcade, J.-C., Krey, V., Kriegler, E., Löschel, A., McCollum, D., Paltsev, S., Rose, S., Shukla, P., Tavoni, M., van der Zwaan, B., and van Vuuren, D. P.: Assessing Transformation Pathways, in: Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Farahani, E., Kadner, S., Seyboth, K., Adler, A., Baum, I., Brunner, S., Eickemeier, P., Kriemann, B., Savolainen, J., Schlömer, S., von Stechow, C., Zwickel, T., Minx, J., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2014.

Coleman, J. S.: Foundations of social theory, The Belknap Press of Harvard University Press, Cambridge, MA, and London, UK, 1994.

Conte, R. and Paolucci, M.: On agent-based modeling and computational social science, Frontiers in Psychology, 5, 668, https://doi.org/10.3389/fpsyg.2014.00668, 2014.

Cooke, I. R., Queenborough, S. A., Mattison, E. H. A., Bailey, A. P., Sandars, D. L., Graves, A. R., Morris, J., Atkinson, P. W., Trawick, P., Freckleton, R. P., Watkinson, A. R., and Sutherland, W. J.: Integrating socio-economics and ecology: A taxonomy of quantitative methods and a review of their use in agro-ecology, J. Appl. Ecol., 46, 269–277, https://doi.org/10.1111/j.1365-2664.2009.01615.x, 2009.

Couzin, I. D.: Collective cognition in animal groups, Trends Cogn. Sci., 13, 36–43, https://doi.org/10.1016/j.tics.2008.10.002, 2009.

Crutzen, P. J.: Geology of mankind, Nature, 415, p. 23, https://doi.org/10.1038/415023a, 2002.

Currarini, S., Marchiori, C., and Tavoni, A.: Network Economics and the Environment: Insights and Perspectives, Environmental and Resource Economics, 65, 159–189, https://doi.org/10.1007/s10640-015-9953-6, 2016.

Dafermos, Y., Nikolaidi, M., and Galanis, G.: A stock-flow-fund ecological macroeconomic model, Ecol. Econ., 131, 191–207, https://doi.org/10.1016/j.ecolecon.2016.08.013, 2017.

Dagum, C.: On the Relationship between Income Inequality measures and social welfare functions, Journal of Econometrics, 43, 91–102, https://doi.org/10.1016/0304-4076(90)90109-7, 1990.

de Marchi, S. and Page, S. E.: Agent-Based Models, Annu. Rev. Polit. Sci., 17, 1–20, https://doi.org/10.1146/annurev-polisci-080812-191558, 2014.

Deadman, P., Robinson, D., Moran, E., and Brondizio, E.: Colonist household decisionmaking and land-use change in the Amazon Rainforest: An agent-based simulation, Environ. Plann. B, 31, 693–709, https://doi.org/10.1068/b3098, 2004.

Deffuant, G., Huet, S., and Amblard, F.: An Individual-Based Model of Innovation Diffusion Mixing Social Value and Individual Benefit, Am. J. Sociol., 110, 1041–1069, https://doi.org/10.1086/430220, 2005.

DeGroot, M. H.: Reaching a Consensus, J. Am. Stat. Assoc., 69, 118–121, https://doi.org/10.1080/01621459.1974.10480137, 1974.

Demirel, G., Vazquez, F., Böhme, G. A., and Gross, T.: Moment-closure approximations for discrete adaptive networks, Physica D, 267, 68–80, https://doi.org/10.1016/j.physd.2013.07.003, 2014.

Dhami, M. K. and Ayton, P.: Bailing and jailing the fast and frugal way, J. Behav. Decis. Making, 14, 141–168, https://doi.org/10.1002/bdm.371, 2001.

Dhami, M. K. and Harries, C.: Fast and frugal versus regression models of human judgement, Think. Reasoning, 7, 5–27, https://doi.org/10.1080/13546780042000019, 2001.

Do, A.-L. and Gross, T.: Contact processes and moment closure on adaptive networks, in: Adaptive Networks: Theory, Models and Applications, edited by: Gross, T. and Sayama, H., Springer and NECSI, Cambridge, Massachusetts, chap. 9, 191–208, https://doi.org/10.1007/978-3-642-01284-6_9, 2009.

Donges, J. F., Lucht, W., Heitzig, J., Cornell, S., Lade, S. J., Schlüter, M., and Barfuss, W.: A taxonomy of co-evolutionary interactions in models of the World-Earth system, in preparation, Earth System Dynamics, 2017a.

Donges, J. F., Lucht, W., Müller-Hansen, F., and Steffen, W.: The technosphere in Earth system analysis: a coevolutionary perspective, The Anthropocene Review, 4, 23–33, https://doi.org/10.1177/2053019616676608, 2017b.

Donges, J. F., Winkelmann, R., Lucht, W., Cornell, S. E., Dyke, J. G., Rockström, J., Heitzig, J., and Schellnhuber, H.-J.: Closing the loop: reconnecting human dynamics to Earth system science, The Anthropocene Review, 4, 151–157, https://doi.org/10.1177/2053019617725537, 2017c.

Durkheim, E.: The rules of sociological method. And selected texts on sociology and its method, The Free Press, New York, 2014.

Edmonds, B. and Meyer, R.: Simulating Social Complexity. A Handbook, Springer, Berlin, New York, https://doi.org/10.1007/978-3-540-93813-2, 2013.

Epstein, J. M.: Agent-based computational models and generative social science, Complexity, 4, 41–60, https://doi.org/10.1002/(SICI)1099-0526(199905/06)4:5<41::AID-CPLX9>3.0.CO;2-F, 1999.

Farmer, J. D., Hepburn, C., Mealy, P., and Teytelboym, A.: A Third Wave in the Economics of Climate Change, Environmental and Resource Economics, 62, 329–357, https://doi.org/10.1007/s10640-015-9965-2, 2015.

Fehr, E. and Fischbacher, U.: The nature of human altruism, Nature, 425, 785–791, https://doi.org/10.1038/nature02043, 2003.

Fehr, E. and Schmidt, K. M.: A Theory of Fairness, Competition, and Cooperation, The Quarterly Journal of Economics, 114, 817–868, https://doi.org/10.1162/003355399556151, 1999.

Feldman, A. M. and Serrano, R.: Welfare economics and social choice theory, 2nd Edn., Springer, New York, 2006.

Feldman, L.: The Opinion Factor: The Effects of Opinionated News on Information Processing and Attitude Change, Polit. Commun., 28, 163–181, https://doi.org/10.1080/10584609.2011.565014, 2011.

Festinger, L., Schachter, S., and Back, K.: Social Pressures in Informal Groups: A Study of Human Factors in Housing, Stanford University Press, Stanford, CA, 1950.

Fischer-Kowalski, M. and Haberl, H.: Tons, joules, and money: Modes of production and their sustainability problems, Soc. Natur. Resour., 10, 61–85, https://doi.org/10.1080/08941929709381009, 1997.

Fishburn, P. C.: Utility Theory, Manage. Sci., 14, 335–378, https://doi.org/10.1287/mnsc.14.5.335, 1968.

Flache, A. and Macy, M. W.: Local Convergence and Global Diversity: From Interpersonal to Social Influence, J. Conflict Resolut., 55, 970–995, https://doi.org/10.1177/0022002711414371, 2011.

Foster, D. and Young, P.: Stochastic evolutionary game dynamics, Theor. Popul. Biol., 38, 219–232, https://doi.org/10.1016/0040-5809(90)90011-J, 1990.

French, J. R. P.: A Formal Theory of Social Power, Psychol. Rev., 63, 181–194, https://doi.org/10.1037/h0046123, 1956.

Friedkin, N. E. and Johnsen, E. C.: Social Influence Network Theory, Cambridge University Press, New York, 2011.

Fudenberg, D. and Levine, D. K.: The Theory of Learning in Games, MIT Press, Cambridge, Massachusetts, 1998.

Garrett, T. J.: Long-run evolution of the global economy – Part 2: Hindcasts of innovation and growth, Earth Syst. Dynam., 6, 673–688, https://doi.org/10.5194/esd-6-673-2015, 2015.

Georgescu-Roegen, N.: The Entropy Law and the Economic Process, Harvard University Press, Cambridge, MA, 1971.

Gibson, C. C., Ostrom, E., and Ahn, T. K.: The concept of scale and the human dimensions of global change: A survey, Ecol. Econ., 32, 217–239, https://doi.org/10.1016/S0921-8009(99)00092-0, 2000.

Gigerenzer, G. and Gaissmaier, W.: Heuristic decision making, Annu. Rev. Psychol., 62, 451–482, https://doi.org/10.1146/annurev-psych-120709-145346, 2011.

Gigerenzer, G. and Selten, R. (Eds.): Bounded rationality: The adaptive toolbox, MIT Press, Cambridge, MA and London, UK, 2002.

Gigerenzer, G. and Todd, P. M.: Simple heuristics that make us smart, Oxford University Press, New York, 1999.

Gigerenzer, G., Hoffrage, U., and Goldstein, D. G.: Fast and frugal heuristics are plausible models of cognition: reply to Dougherty, Franco-Watkins, and Thomas (2008), Psychol. Rev., 115, 230–239, https://doi.org/10.1037/0033-295X.115.1.230, 2008.

Gilbert, N.: Agent-based models, Sage, Thousand Oaks, CA, USA, 2008.

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

Gillespie, C.: Moment-closure approximations for mass-action models, IET Syst. Biol., 3, 52–58, https://doi.org/10.1049/iet-syb:20070031, 2009.

Gintis, H.: The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences, Princeton University Press, Princeton and Oxford, 2009.

Godley, W. and Lavoie, M.: Monetary Economics. An Integrated Approach to Credit, Money, Income, Production and Wealth, Palgrave Macmillan, New York, 2007.

Goodman, L. A.: Population Growth of the Sexes, Biometrics, 9, 212–225, https://doi.org/10.2307/3001852, 1953.

Grimm, V. and Railsback, S. F.: Individual-based modeling and ecology, Princeton Series in Theoretical and Computational Biology, Princeton University Press, Princeton, NJ, 2005.

Groeneveld, J., Müller, B., Buchmann, C., Dressler, G., Guo, C., Hase, N., Hoffmann, F., John, F., Klassert, C., Lauf, T., Liebelt, V., Nolzen, H., Pannicke, N., Schulze, J., Weise, H., and Schwarz, N.: Theoretical foundations of human decision-making in agent-based land use models – A review, Environ. Modell. Softw., 87, 39–48, https://doi.org/10.1016/j.envsoft.2016.10.008, 2017.

Gross, T. and Blasius, B.: Adaptive coevolutionary networks: a review, J. R. Soc. Interface, 5, 259–271, https://doi.org/10.1098/rsif.2007.1229, 2008.

Gross, T., D'Lima, C. J. D., and Blasius, B.: Epidemic dynamics on an adaptive network, Phys. Rev. Lett., 96, 208701, https://doi.org/10.1103/PhysRevLett.96.208701, 2006.

Hamill, L. and Gilbert, N.: Agent-Based Modelling in Economics, Wiley, Chichester, UK, 2016.

Hansson, S. O.: Social Choice with procedural preferences, Soc. Choice Welfare, 13, 215–230, https://doi.org/10.1007/BF00183352, 1996.

Hardt, L. and O'Neill, D. W.: Ecological Macroeconomic Models: Assessing Current Developments, Ecol. Econ., 134, 198–211, https://doi.org/10.1016/j.ecolecon.2016.12.027, 2017.

Harsanyi, J. C. and Selten, R.: A General Theory of Equilibrium Selection in Games, MIT Press, Cambridge, MA, 1988.

Hauser, J. R., Ding, M., and Gaskin, S. P.: Non-compensatory (and Compensatory) Models of Consideration-Set Decisions, in: Proceedings of the Sawtooth Software Conference, May, 2009.

Havlík, P., Schneider, U. A., Schmid, E., Böttcher, H., Fritz, S., Skalský, R., Aoki, K., Cara, S. D., Kindermann, G., Kraxner, F., Leduc, S., McCallum, I., Mosnier, A., Sauer, T., and Obersteiner, M.: Global land-use implications of first and second generation biofuel targets, Energ. Policy, 39, 5690–5702, https://doi.org/10.1016/j.enpol.2010.03.030, 2011.

Heckbert, S.: MayaSim: An Agent-Based Model of the Ancient Maya Social-Ecological System, JASSS-J. Artif. Soc. S., 16, 11, https://doi.org/10.18564/jasss.2305, 2013.

Heckbert, S., Baynes, T., and Reeson, A.: Agent-based modeling in ecological economics, Ann. NY Acad. Sci., 1185, 39–53, https://doi.org/10.1111/j.1749-6632.2009.05286.x, 2010.

Hedström, P.: Dissecting the social: On the principles of analytical sociology, Cambridge University Press, Cambridge, UK, 2005.

Hedström, P. and Udehn, L.: Analytical Sociology and Theories of the Middle Range, in: The Oxford Handbook of Analytical Sociology, edited by: Hedström, P. and Bearman, P., Oxford University Press, Oxford and New York, chap. 2, 25–47, 2009.

Hedström, P. and Ylikoski, P.: Causal Mechanisms in the Social Sciences, Annu. Rev. Sociol., 36, 49–67, https://doi.org/10.1146/annurev.soc.012809.102632, 2010.

Hegselmann, R. and Krause, U.: Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation, JASSS-J. Artif. Soc. S., 5, 1–33, 2002.

Heitzig, J.: Bottom-Up Strategic Linking of Carbon Markets: Which Climate Coalitions Would Farsighted Players Form?, FEEM Working Paper No. 48.2013, https://doi.org/10.2139/ssrn.2274724, 2013.

Heitzig, J. and Simmons, F. W.: Some chance for consensus: voting methods for which consensus is an equilibrium, Soc. Choice Welfare, 38, 43–57, https://doi.org/10.1007/s00355-010-0517-y, 2012.

Heitzig, J., Lessmann, K., and Zou, Y.: Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games, P. Natl. Acad. Sci. USA, 108, 15739–15744, https://doi.org/10.1073/pnas.1106265108, 2011.

Hertel, T. W., Rose, S. K., and Tol, R. S. J. (Eds.): Economic analysis of land use in global climate change policy, Routledge, London and New York, 2009.

Hertwig, R. and Herzog, S. M.: Fast and Frugal Heuristics: Tools of Social Rationality, Soc. Cognition, 27, 661–698, https://doi.org/10.1521/soco.2009.27.5.661, 2009.

Hilbert, M.: Toward a Synthesis of Cognitive Biases: How Noisy Information Processing Can Bias Human Decision Making, Psychol. Bull., 138, 211–237, https://doi.org/10.1037/a0025940, 2012.

Hodgson, G. M.: Meanings of methodological individualism, Journal of Economic Methodology, 14, 211–226, https://doi.org/10.1080/13501780701394094, 2007.

Holme, P. and Newman, M. E. J.: Nonequilibrium phase transition in the coevolution of networks and opinions, Phys. Rev. E, 74, 056108, https://doi.org/10.1103/PhysRevE.74.056108, 2006.

Holme, P. and Saramäki, J.: Temporal networks, Phys. Rep., 519, 97–125, https://doi.org/10.1016/j.physrep.2012.03.001, 2012.

Homans, G. C.: The human group, Harcourt, Brace & World, New York, 1950.

Huckfeldt, R., Johnson, P. E., and Sprague, J.: Political Disagreement. The Survival of Diverse Opinions within Communication Networks, Cambridge University Press, Cambridge, UK, 2004.

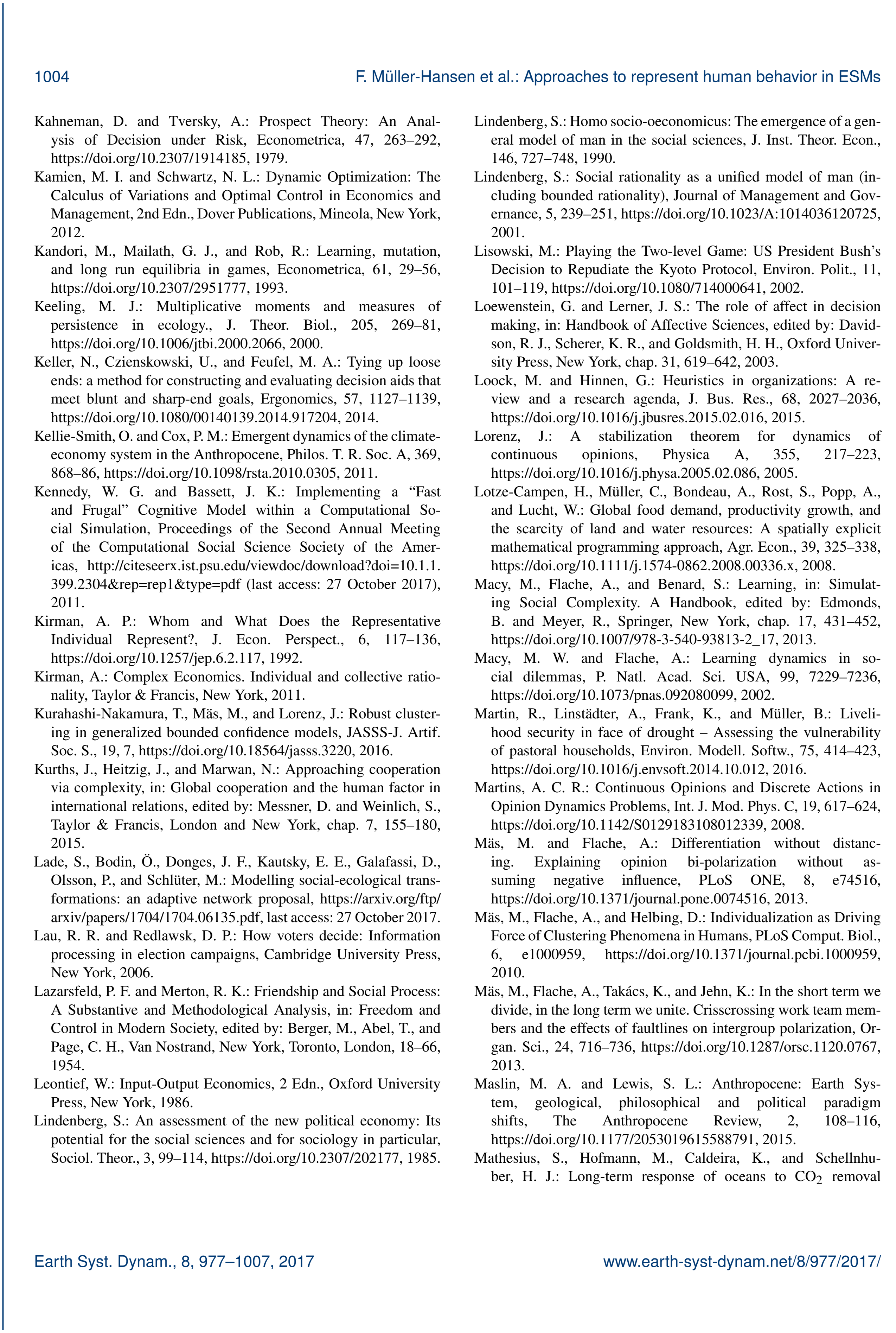
Hughes, B.: International Futures: Choices in the face of uncertainty, 3rd Edn., Westview Press, Boulder, CO, 1999.

IPCC: Climate Change 2014: Synthesis Report. Contribution of Working Group I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, IPCC, Geneva, 2014.

Jamison, D. T. and Jamison, J.: Characterizing the Amount and Speed of Discounting Procedures, Journal of Benefit-Cost Analysis, 2, 1, https://doi.org/10.2202/2152-2812.1031, 2011.

Jarvis, A. J., Jarvis, S. J., and Hewitt, C. N.: Resource acquisition, distribution and end-use efficiencies and the growth of industrial society, Earth Syst. Dynam., 6, 689–702, https://doi.org/10.5194/esd-6-689-2015, 2015.

Jones, D. A.: The polarizing effect of new media messages, International Journal of Public Opinion Research, 14, 158–174, https://doi.org/10.1093/Ijpor/14.2.158, 2002.



Kahneman, D. and Tversky, A.: Prospect Theory: An Analysis of Decision under Risk, Econometrica, 47, 263–292, https://doi.org/10.2307/1914185, 1979.

Kamien, M. I. and Schwartz, N. L.: Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management, 2nd Edn., Dover Publications, Mineola, New York, 2012.

Kandori, M., Mailath, G. J., and Rob, R.: Learning, mutation, and long run equilibria in games, Econometrica, 61, 29–56, https://doi.org/10.2307/2951777, 1993.

Keeling, M. J.: Multiplicative moments and measures of persistence in ecology., J. Theor. Biol., 205, 269–81, https://doi.org/10.1006/jtbi.2000.2066, 2000.

Keller, N., Czienskowski, U., and Feufel, M. A.: Tying up loose ends: a method for constructing and evaluating decision aids that meet blunt and sharp-end goals, Ergonomics, 57, 1127–1139, https://doi.org/10.1080/00140139.2014.917204, 2014.

Kellie-Smith, O. and Cox, P. M.: Emergent dynamics of the climate-economy system in the Anthropocene, Philos. T. R. Soc. A, 369, 868–86, https://doi.org/10.1098/rsta.2010.0305, 2011.

Kennedy, W. G. and Bassett, J. K.: Implementing a "Fast and Frugal" Cognitive Model within a Computational Social Simulation, Proceedings of the Second Annual Meeting of the Computational Social Science Society of the Americas, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.399.2304&rep=rep1&type=pdf (last access: 27 October 2017), 2011.

Kirman, A. P.: Whom and What Does the Representative Individual Represent?, J. Econ. Perspect., 6, 117–136, https://doi.org/10.1257/jep.6.2.117, 1992.

Kirman, A.: Complex Economics. Individual and collective rationality, Taylor & Francis, New York, 2011.

Kurahashi-Nakamura, T., Mäs, M., and Lorenz, J.: Robust clustering in generalized bounded confidence models, JASSS-J. Artif. Soc. S., 19, 7, https://doi.org/10.18564/jasss.3220, 2016.

Kurths, J., Heitzig, J., and Marwan, N.: Approaching cooperation via complexity, in: Global cooperation and the human factor in international relations, edited by: Messner, D. and Weinlich, S., Taylor & Francis, London and New York, chap. 7, 155–180, 2015.

Lade, S., Bodin, Ö., Donges, J. F., Kautsky, E. E., Galafassi, D., Olsson, P., and Schlüter, M.: Modelling social-ecological transformations: an adaptive network proposal, https://arxiv.org/ftp/arxiv/papers/1704/1704.06135.pdf, last access: 27 October 2017.

Lau, R. R. and Redlawsk, D. P.: How voters decide: Information processing in election campaigns, Cambridge University Press, New York, 2006.

Lazarsfeld, P. F. and Merton, R. K.: Friendship and Social Process: A Substantive and Methodological Analysis, in: Freedom and Control in Modern Society, edited by: Berger, M., Abel, T., and Page, C. H., Van Nostrand, New York, Toronto, London, 18–66, 1954.

Leontief, W.: Input-Output Economics, 2 Edn., Oxford University Press, New York, 1986.

Lindenberg, S.: An assessment of the new political economy: Its potential for the social sciences and for sociology in particular, Sociol. Theor., 3, 99–114, https://doi.org/10.2307/202177, 1985.

Lindenberg, S.: Homo socio-oeconomicus: The emergence of a general model of man in the social sciences, J. Inst. Theor. Econ., 146, 727–748, 1990.

Lindenberg, S.: Social rationality as a unified model of man (including bounded rationality), Journal of Management and Governance, 5, 239–251, https://doi.org/10.1023/A:1014036120725, 2001.

Lisowski, M.: Playing the Two-level Game: US President Bush's Decision to Repudiate the Kyoto Protocol, Environ. Polit., 11, 101–119, https://doi.org/10.1080/714000641, 2002.

Loewenstein, G. and Lerner, J. S.: The role of affect in decision making, in: Handbook of Affective Sciences, edited by: Davidson, R. J., Scherer, K. R., and Goldsmith, H. H., Oxford University Press, New York, chap. 31, 619–642, 2003.

Loock, M. and Hinnen, G.: Heuristics in organizations: A review and a research agenda, J. Bus. Res., 68, 2027–2036, https://doi.org/10.1016/j.jbusres.2015.02.016, 2015.

Lorenz, J.: A stabilization theorem for dynamics of continuous opinions, Physica A, 355, 217–223, https://doi.org/10.1016/j.physa.2005.02.086, 2005.

Lotze-Campen, H., Müller, C., Bondeau, A., Rost, S., Popp, A., and Lucht, W.: Global food demand, productivity growth, and the scarcity of land and water resources: A spatially explicit mathematical programming approach, Agr. Econ., 39, 325–338, https://doi.org/10.1111/j.1574-0862.2008.00336.x, 2008.

Macy, M., Flache, A., and Benard, S.: Learning, in: Simulating Social Complexity. A Handbook, edited by: Edmonds, B. and Meyer, R., Springer, New York, chap. 17, 431–452, https://doi.org/10.1007/978-3-540-93813-2_17, 2013.

Macy, M. W. and Flache, A.: Learning dynamics in social dilemmas, P. Natl. Acad. Sci. USA, 99, 7229–7236, https://doi.org/10.1073/pnas.092080099, 2002.

Martin, R., Linstädter, A., Frank, K., and Müller, B.: Livelihood security in face of drought – Assessing the vulnerability of pastoral households, Environ. Modell. Softw., 75, 414–423, https://doi.org/10.1016/j.envsoft.2014.10.012, 2016.

Martins, A. C. R.: Continuous Opinions and Discrete Actions in Opinion Dynamics Problems, Int. J. Mod. Phys. C, 19, 617–624, https://doi.org/10.1142/S0129183108012339, 2008.

Mäs, M. and Flache, A.: Differentiation without distancing. Explaining opinion bi-polarization without assuming negative influence, PLoS ONE, 8, e74516, https://doi.org/10.1371/journal.pone.0074516, 2013.

Mäs, M., Flache, A., and Helbing, D.: Individualization as Driving Force of Clustering Phenomena in Humans, PLoS Comput. Biol., 6, e1000959, https://doi.org/10.1371/journal.pcbi.1000959, 2010.

Mäs, M., Flache, A., Takács, K., and Jehn, K.: In the short term we divide, in the long term we unite. Crisscrossing work team members and the effects of faultlines on intergroup polarization, Organ. Sci., 24, 716–736, https://doi.org/10.1287/orsc.1120.0767, 2013.

Maslin, M. A. and Lewis, S. L.: Anthropocene: Earth System, geological, philosophical and political paradigm shifts, The Anthropocene Review, 2, 108–116, https://doi.org/10.1177/2053019615588791, 2015.

Mathesius, S., Hofmann, M., Caldeira, K., and Schellnhuber, H. J.: Long-term response of oceans to CO_2 removal

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

from the atmosphere, Nature Climate Change, 5, 1107–1113, https://doi.org/10.1038/nclimate2729, 2015.

Matthews, R. B., Gilbert, N. G., Roach, A., Polhill, J. G., and Gotts, N. M.: Agent-based land-use models: a review of applications, Landscape Ecol., 22, 1447–1459, https://doi.org/10.1007/s10980-007-9135-1, 2007.

McPherson, M., Smith-Lovin, L., and Cook, J. M.: Birds of a Feather: Homophily in Social Networks, Annu. Rev. Sociol., 27, 415–444, https://doi.org/10.1146/annurev.soc.27.1.415, 2001.

Meadows, D. H., Meadows, D. L., Randers, J., and Behrens, W. W.: The Limits to Growth, Universe Books, New York, 1972.

Meadows, D. H., Randers, J., and Meadows, D. L.: Limits to Growth: The 30-Year Update, Earthscan, London, 2004.

Merton, R. K.: Social Theory and Social Structure, The Free Press, New York, 1957.

Meyfroidt, P.: Environmental cognitions, land change, and social–ecological feedbacks: an overview, Journal of Land Use Science, 8, 341–367, https://doi.org/10.1080/1747423X.2012.667452, 2013.

Michetti, M.: Modelling Land Use, Land-Use Change, and Forestry in Climate Change: A Review of Major Approaches, FEEM Working Paper No. 46.2012, https://doi.org/10.2139/ssrn.2122298, 2012.

Miller, R. E. and Blair, P. D.: Input-Output Analysis: Foundations and Extensions, 2nd Edn., Cambridge University Press, Cambridge, UK, 2009.

Millner, A.: On welfare frameworks and catastrophic climate risks, J. Environ. Econ. Manag., 65, 310–325, https://doi.org/10.1016/j.jeem.2012.09.006, 2013.

Moss, R., Edmonds, J., Hibbard, K., Manning, M., Rose, S., Van Vuuren, D., Carter, T., Emori, S., Kainuma, M., Kram, T., Meehl, G., Mitchell, J., Nakicenovic, N., Riahi, K., Smith, S., Stouffer, R., Thomson, A., Weyant, J., and Wilbanks, T.: The next generation of scenarios for climate change research and assessment, Nature, 463, 747–756, https://doi.org/10.1038/nature08823, 2010.

Mueller, D. C.: Public Choice III, Cambridge University Press, Cambridge, UK, 2003.

Murray-Rust, D., Brown, C., van Vliet, J., Alam, S. J., Robinson, D. T., Verburg, P. H., and Rounsevell, M.: Combining agent functional types, capitals and services to model land use dynamics, Environ. Modell. Softw., 59, 187–201, https://doi.org/10.1016/j.envsoft.2014.05.019, 2014.

Myers, D. G.: Polarizing Effects of Social Interaction, in: Group Decision Making, edited by: Brandstätter, H., Davis, J. H., and Stocker-Kreichgauer, G., Academic Press, London, chap. 6, 125–161, 1982.

Nowak, A., Szamrej, J., and Latané, B.: From Private Attitude to Public Opinion: A Dynamic Theory of Social Impact, Psychol. Rev., 97, 362–376, https://doi.org/10.1037/0033-295X.97.3.362, 1990.

Nowak, M. A.: Evolutionary Dynamics – Exploring the Equations of Life, The Belknap Press of Harvad University Press, Cambridge, MA and London, UK, 2006.

Opp, K.-D.: Contending conceptions of the theory of rational action, J. Theor. Polit., 11, 171–202, https://doi.org/10.1177/0951692899011002002, 1999.

Ordeshook, P. C.: Game theory and political theory. An Introduction, Cambridge University Press, Cambridge, UK, New York and Melbourne, 1986.

Ostrom, E.: Governing the Commons: The Evolution of Institutions for Collective Action, Cambridge University Press, Cambridge, UK, 1990.

Page, S. E.: What Sociologists Should Know About Complexity, Ann. Rev. Sociol., 41, 21–41, https://doi.org/10.1146/annurev-soc-073014-112230, 2015.

Palmer, P. I. and Smith, M. J.: Model human adaptation to climate change, Nature, 512, 365–366, https://doi.org/10.1038/512365a, 2014.

Parker, D. C., Manson, S. M., Janssen, M. a., Hoffmann, M. J., and Deadman, P.: Multi-Agent Systems for the Simulation of Land-Use and Land-Cover Change: A Review, Ann. Assoc. Am. Geogr., 93, 314–337, https://doi.org/10.1111/1467-8306.9302004, 2003.

Perc, M. and Szolnoki, A.: Coevolutionary games-A mini review, BioSystems, 99, 109–125, https://doi.org/10.1016/j.biosystems.2009.10.003, 2010.

Perman, R., Ma, Y., McGilvray, J., and Common, M.: Natural resource and environmental economics, 3rd Edn., Pearson Education, Essex, UK, 2003.

Pineda, M., Toral, R., and Hernández-García, E.: Noisy continuous-opinion dynamics, J. Stat. Mech.-Theory E., 08, P08001, https://doi.org/10.1088/1742-5468/2009/08/P08001, 2009.

Puga, J. L., Krzywinski, M., and Altman, N.: Points of Significance: Bayes' theorem, Nat. Methods, 12, 277–278, https://doi.org/10.1038/nmeth.3335, 2015.

Putnam, R. D.: Diplomacy and domestic politics: the logic of two-level games, Int. Organ., 42, 427–460, https://doi.org/10.1017/S0020818300027697, 1988.

Rabin, M.: A perspective on psychology and economics, Eur. Econ. Rev., 46, 657–685, https://doi.org/10.1016/S0014-2921(01)00207-0, 2002.

Reif, F.: Fundamentals of statistical and thermal physics, McGraw-Hill, New York, 1965.

Rizvi, S. A. T.: The microfoundations project in general equilibrium theory, Cambridge J. Econ., 18, 357–377, https://doi.org/10.1093/oxfordjournals.cje.a035280, 1994.

Rogers, T., Clifford-Brown, W., Mills, C., and Galla, T.: Stochastic oscillations of adaptive networks: application to epidemic modelling, J. Stat. Mech.-Theory E, 2012, P08018, https://doi.org/10.1088/1742-5468/2012/08/P08018, 2012.

Romer, P. M.: Increasing Returns and Long-Run Growth, J. Polit. Econ., 94, 1002–1037, https://doi.org/10.1086/261420, 1986.

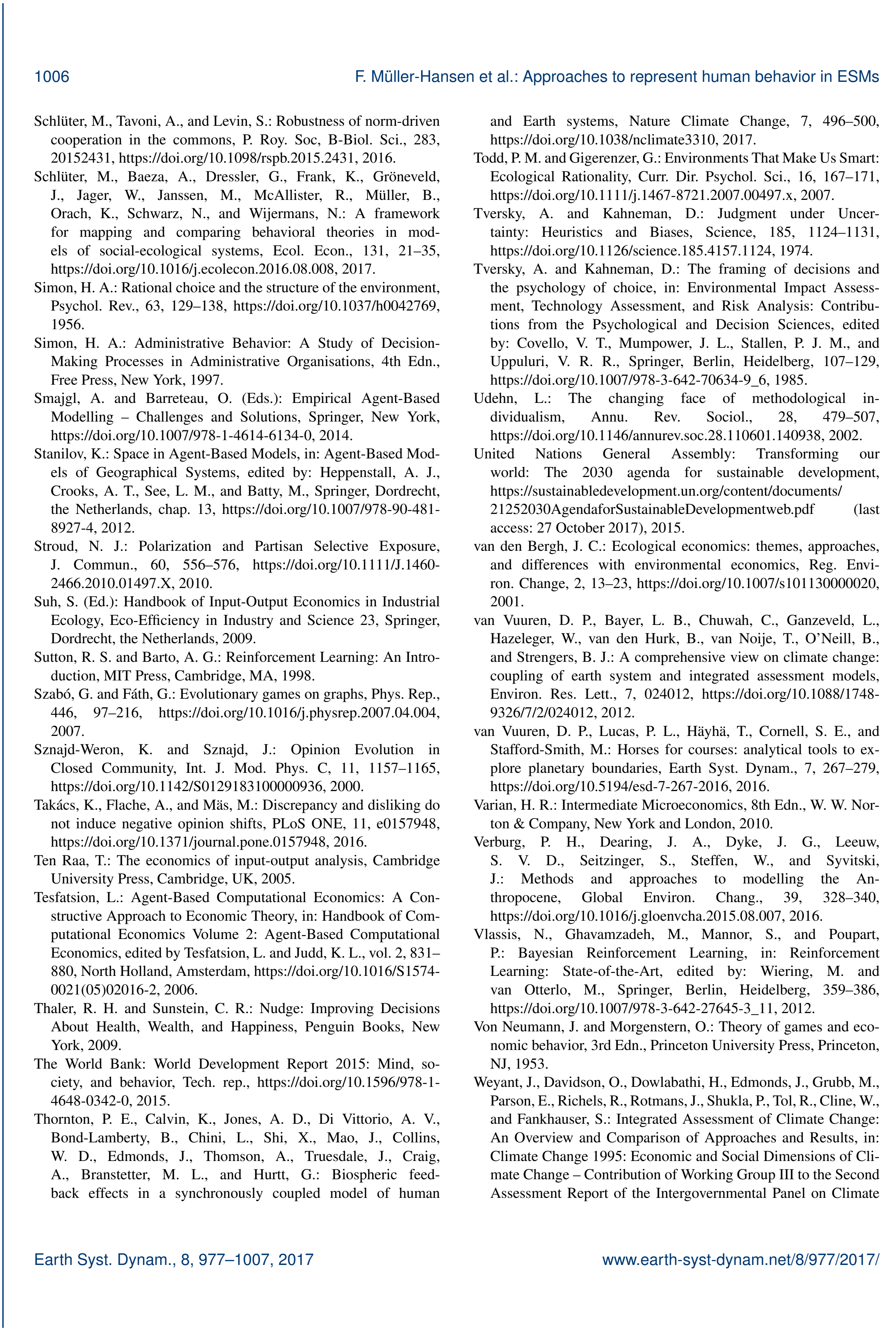
Rosenberg, A.: Philosophy of Social Science, 4th Edn., Westview Press, Boulder, CO, 2012.

Schelling, T. C.: Micromotives and Macrobehavior, W. W. Norton and Company, New York, 1978.

Schleussner, C.-F., Donges, J. F., Engemann, D. A., and Levermann, A.: Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure, Nature Scientific Reports, 6, 30790, https://doi.org/10.1038/srep30790, 2016.

Schlüter, M., Mcallister, R. R. J., Arlinghaus, R., Bunnefeld, N., Eisenack, K., Hölker, F., Milner-Gulland, E. J., Müller, B., Nicholson, E., Quaas, M., and Stöven, M.: New Horizons for Managing the Environment: A Review of Coupled Social-Ecological Systems Modeling, Nat. Resour. Model., 25, 219–272, https://doi.org/10.1111/j.1939-7445.2011.00108.x, 2012.

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).



Schlüter, M., Tavoni, A., and Levin, S.: Robustness of norm-driven cooperation in the commons, P. Roy. Soc, B-Biol. Sci., 283, 20152431, https://doi.org/10.1098/rspb.2015.2431, 2016.

Schlüter, M., Baeza, A., Dressler, G., Frank, K., Gröneveld, J., Jager, W., Janssen, M., McAllister, R., Müller, B., Orach, K., Schwarz, N., and Wijermans, N.: A framework for mapping and comparing behavioral theories in models of social-ecological systems, Ecol. Econ., 131, 21–35, https://doi.org/10.1016/j.ecolecon.2016.08.008, 2017.

Simon, H. A.: Rational choice and the structure of the environment, Psychol. Rev., 63, 129–138, https://doi.org/10.1037/h0042769, 1956.

Simon, H. A.: Administrative Behavior: A Study of Decision-Making Processes in Administrative Organisations, 4th Edn., Free Press, New York, 1997.

Smajgl, A. and Barreteau, O. (Eds.): Empirical Agent-Based Modelling – Challenges and Solutions, Springer, New York, https://doi.org/10.1007/978-1-4614-6134-0, 2014.

Stanilov, K.: Space in Agent-Based Models, in: Agent-Based Models of Geographical Systems, edited by: Heppenstall, A. J., Crooks, A. T., See, L. M., and Batty, M., Springer, Dordrecht, the Netherlands, chap. 13, https://doi.org/10.1007/978-90-481-8927-4, 2012.

Stroud, N. J.: Polarization and Partisan Selective Exposure, J. Commun., 60, 556–576, https://doi.org/10.1111/J.1460-2466.2010.01497.X, 2010.

Suh, S. (Ed.): Handbook of Input-Output Economics in Industrial Ecology, Eco-Efficiency in Industry and Science 23, Springer, Dordrecht, the Netherlands, 2009.

Sutton, R. S. and Barto, A. G.: Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998.

Szabó, G. and Fáth, G.: Evolutionary games on graphs, Phys. Rep., 446, 97–216, https://doi.org/10.1016/j.physrep.2007.04.004, 2007.

Sznajd-Weron, K. and Sznajd, J.: Opinion Evolution in Closed Community, Int. J. Mod. Phys. C, 11, 1157–1165, https://doi.org/10.1142/S0129183100000936, 2000.

Takács, K., Flache, A., and Mäs, M.: Discrepancy and disliking do not induce negative opinion shifts, PLoS ONE, 11, e0157948, https://doi.org/10.1371/journal.pone.0157948, 2016.

Ten Raa, T.: The economics of input-output analysis, Cambridge University Press, Cambridge, UK, 2005.

Tesfatsion, L.: Agent-Based Computational Economics: A Constructive Approach to Economic Theory, in: Handbook of Computational Economics Volume 2: Agent-Based Computational Economics, edited by Tesfatsion, L. and Judd, K. L., vol. 2, 831–880, North Holland, Amsterdam, https://doi.org/10.1016/S1574-0021(05)02016-2, 2006.

Thaler, R. H. and Sunstein, C. R.: Nudge: Improving Decisions About Health, Wealth, and Happiness, Penguin Books, New York, 2009.

The World Bank: World Development Report 2015: Mind, society, and behavior, Tech. rep., https://doi.org/10.1596/978-1-4648-0342-0, 2015.

Thornton, P. E., Calvin, K., Jones, A. D., Di Vittorio, A. V., Bond-Lamberty, B., Chini, L., Shi, X., Mao, J., Collins, W. D., Edmonds, J., Thomson, A., Truesdale, J., Craig, A., Branstetter, M. L., and Hurtt, G.: Biospheric feedback effects in a synchronously coupled model of human and Earth systems, Nature Climate Change, 7, 496–500, https://doi.org/10.1038/nclimate3310, 2017.

Todd, P. M. and Gigerenzer, G.: Environments That Make Us Smart: Ecological Rationality, Curr. Dir. Psychol. Sci., 16, 167–171, https://doi.org/10.1111/j.1467-8721.2007.00497.x, 2007.

Tversky, A. and Kahneman, D.: Judgment under Uncertainty: Heuristics and Biases, Science, 185, 1124–1131, https://doi.org/10.1126/science.185.4157.1124, 1974.

Tversky, A. and Kahneman, D.: The framing of decisions and the psychology of choice, in: Environmental Impact Assessment, Technology Assessment, and Risk Analysis: Contributions from the Psychological and Decision Sciences, edited by: Covello, V. T., Mumpower, J. L., Stallen, P. J. M., and Uppuluri, V. R. R., Springer, Berlin, Heidelberg, 107–129, https://doi.org/10.1007/978-3-642-70634-9_6, 1985.

Udehn, L.: The changing face of methodological individualism, Annu. Rev. Sociol., 28, 479–507, https://doi.org/10.1146/annurev.soc.28.110601.140938, 2002.

United Nations General Assembly: Transforming our world: The 2030 agenda for sustainable development, https://sustainabledevelopment.un.org/content/documents/21252030AgendaforSustainableDevelopmentweb.pdf (last access: 27 October 2017), 2015.

van den Bergh, J. C.: Ecological economics: themes, approaches, and differences with environmental economics, Reg. Environ. Change, 2, 13–23, https://doi.org/10.1007/s101130000020, 2001.

van Vuuren, D. P., Bayer, L. B., Chuwah, C., Ganzeveld, L., Hazeleger, W., van den Hurk, B., van Noije, T., O'Neill, B., and Strengers, B. J.: A comprehensive view on climate change: coupling of earth system and integrated assessment models, Environ. Res. Lett., 7, 024012, https://doi.org/10.1088/1748-9326/7/2/024012, 2012.

van Vuuren, D. P., Lucas, P. L., Häyhä, T., Cornell, S. E., and Stafford-Smith, M.: Horses for courses: analytical tools to explore planetary boundaries, Earth Syst. Dynam., 7, 267–279, https://doi.org/10.5194/esd-7-267-2016, 2016.

Varian, H. R.: Intermediate Microeconomics, 8th Edn., W. W. Norton & Company, New York and London, 2010.

Verburg, P. H., Dearing, J. A., Dyke, J. G., Leeuw, S. V. D., Seitzinger, S., Steffen, W., and Syvitski, J.: Methods and approaches to modelling the Anthropocene, Global Environ. Chang., 39, 328–340, https://doi.org/10.1016/j.gloenvcha.2015.08.007, 2016.

Vlassis, N., Ghavamzadeh, M., Mannor, S., and Poupart, P.: Bayesian Reinforcement Learning, in: Reinforcement Learning: State-of-the-Art, edited by: Wiering, M. and van Otterlo, M., Springer, Berlin, Heidelberg, 359–386, https://doi.org/10.1007/978-3-642-27645-3_11, 2012.

Von Neumann, J. and Morgenstern, O.: Theory of games and economic behavior, 3rd Edn., Princeton University Press, Princeton, NJ, 1953.

Weyant, J., Davidson, O., Dowlabathi, H., Edmonds, J., Grubb, M., Parson, E., Richels, R., Rotmans, J., Shukla, P., Tol, R., Cline, W., and Fankhauser, S.: Integrated Assessment of Climate Change: An Overview and Comparison of Approaches and Results, in: Climate Change 1995: Economic and Social Dimensions of Climate Change – Contribution of Working Group III to the Second Assessment Report of the Intergovernmental Panel on Climate

Reproduced from: F. Müller-Hansen et al., Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches, Earth Syst. Dyn., vol. 8, no. 4, pp. 977-1007, 2017, doi: 10.5194/esd-8-977-2017. Published under Creative Commons Attribution License 3.0 (CC BY).

Change, edited by: Bruce, J. P., Lee, H., and Haites, E. F., Cambridge University Press, Cambridge, UK, New York and Melbourne, 367–396, 1996.

Wickens, M.: Macroeconomic Theory. A Dynamic General Equilibrium Approach, Princeton University Press, Princeton and Oxford, 2008.

Wiedermann, M., Donges, J. F., Heitzig, J., Lucht, W., and Kurths, J.: Macroscopic description of complex adaptive networks coevolving with dynamic node states, Phys. Rev. E, 91, 052801, https://doi.org/10.1103/PhysRevE.91.052801, 2015.

Wiedmann, T.: A review of recent multi-region input-output models used for consumption-based emission and resource accounting, Ecol. Econ., 69, 211–222, https://doi.org/10.1016/j.ecolecon.2009.08.026, 2009.

Wimmer, A. and Lewis, K.: Beyond and Below Racial Homophily: ERG Models of a Friendship Network Documented on Facebook, Am. J. Sociol., 116, 583–642, https://doi.org/10.1086/653658, 2010.

Wood, W.: Attitude Change: Persuasion and Social Influence, Annu. Rev. Psychol., 51, 539–570, https://doi.org/10.1146/annurev.psych.51.1.539, 2000.

Taxonomies for structuring models for World-Earth systems analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops

Jonathan F. Donges[1,2], Wolfgang Lucht[1,3,4], Sarah E. Cornell[2], Jobst Heitzig[5], Wolfram Barfuss[1,6], Steven J. Lade[2,7,8], and Maja Schlüter[2]

[1]Earth System Analysis, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Telegrafenberg A31, 14473 Potsdam, Germany
[2]Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden
[3]Department of Geography, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany
[4]Integrative Research Institute on Transformations of Human-Environment Systems, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany
[5]Complexity Science, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Telegrafenberg A31, 14473 Potsdam, Germany
[6]Department of Physics, Humboldt University, Newtonstr. 15, 12489 Berlin, Germany
[7]Fenner School of Environment and Society, The Australian National University, Building 141, Linnaeus way, Canberra, Australian Capital Territory, 2601, Australia
[8]Bolin Centre for Climate Research, Stockholm University, Stockholm, Sweden

Correspondence to: Jonathan F. Donges (donges@pik-potsdam.de)

Abstract.

In the Anthropocene, the social dynamics of human societies have become critical to understanding planetary-scale Earth system dynamics. The conceptual foundations of Earth system modelling have externalised social processes in ways that now hinder progress in understanding Earth resilience and informing governance of global environmental change. New approaches to global modelling of the human World are needed to address these challenges. The current modelling landscape is highly diverse and heterogeneous, ranging from purely biophysical Earth System Models, to hybrid macro-economic Integrated Assessments Models, to a plethora of models of socio-cultural dynamics. World-Earth models capable of simulating complex and entangled human-Earth system processes of the Anthropocene are currently not available. They will need to draw on and selectively integrate elements from the diverse range of fields and approaches, so future World-Earth modellers require a structured approach to identify, classify, select, combine and critique model components from multiple modeling traditions. Here, we develop taxonomies for ordering the multitude of societal and biophysical subsystems and their interactions. We suggest three taxa for modelled subsystems: (i) biophysical, where dynamics is usually represented by “natural laws" of physics, chemistry or ecology (i.e., the usual components of Earth system models), (ii) socio-cultural, dominated by processes of human behaviour, decision making and collective social dynamics (e.g., politics, institutions, social networks, and even science itself), and (iii) socio-metabolic, dealing with the material interactions of social and biophysical subsystems (e.g., human bodies, natural resources and agriculture). We show how higher-order taxonomies can be derived for classifying and describing the interactions between two or more subsystems. This then allows us to highlight the kinds of social-ecological feedback loops where new

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

modelling efforts need to be directed. As an example, we apply the taxonomy to a stylised World-Earth system model that endogenises socially transmitted choice of discount rates in a greenhouse gas emissions game to illustrate the effects of social-ecological feedback loops that are usually not considered in current modelling efforts. The proposed taxonomy can contribute to guiding the design and operational development of more comprehensive World-Earth models for understanding Earth resilience and charting sustainability transitions within planetary boundaries and other future trajectories in the Anthropocene.

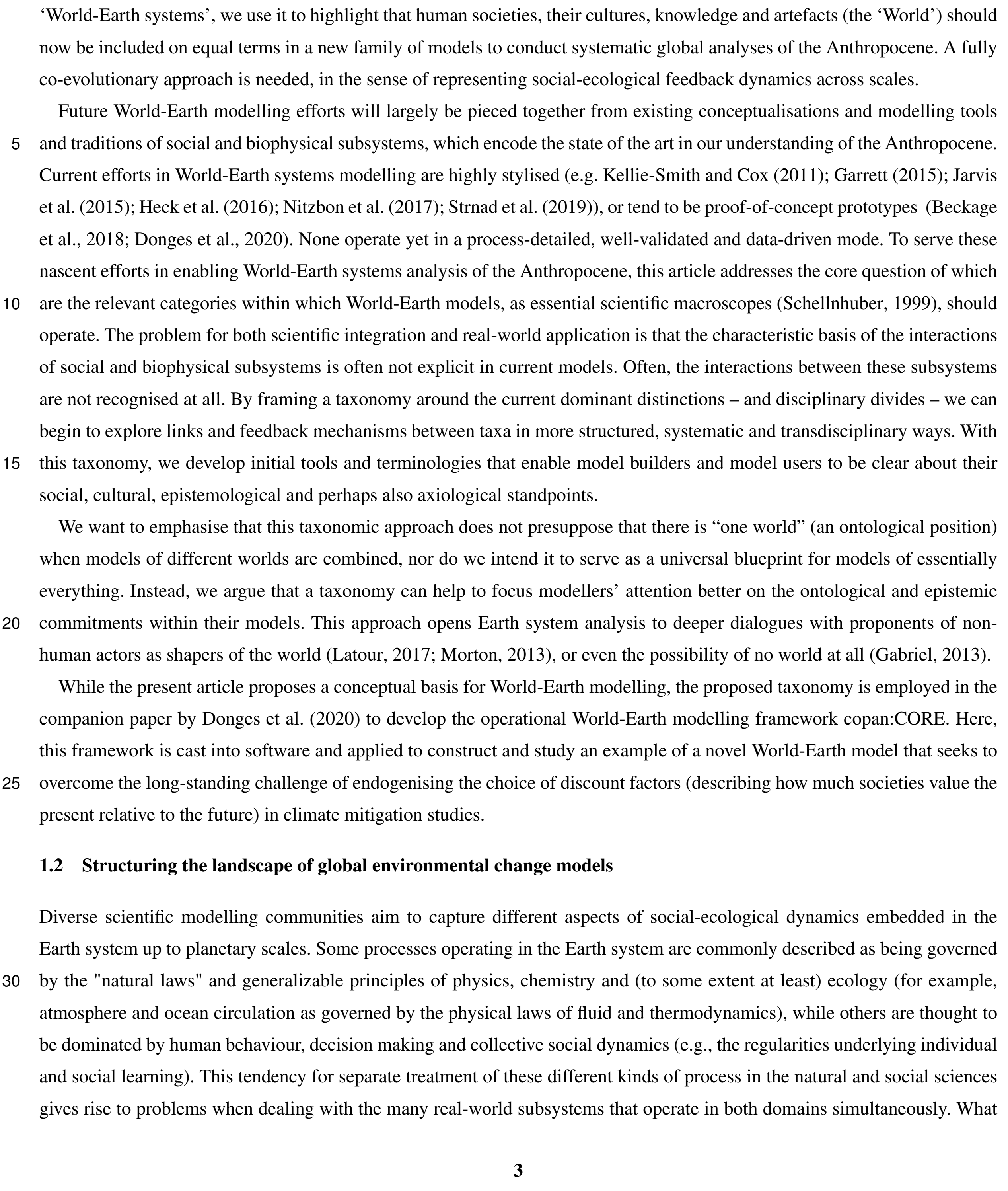
1 Introduction

1.1 Revisiting Earth system analysis for the Anthropocene

In the age of the Anthropocene, human societies have emerged as a planetary-scale geological force shaping the future trajectory of the whole Earth system (Crutzen, 2002; Steffen et al., 2007; Lewis and Maslin, 2015; Waters et al., 2016; Lenton and Latour, 2018; Steffen et al., 2018). Cumulative greenhouse gas emissions and extensive modifications of the biosphere have accelerated since the neolithic and industrial revolutions, especially through the rapid globalisation of social-economic systems during the 20th century, threatening the stability of the interglacial state (Lenton et al., 2016) that has enabled the development and wellbeing of human societies (Rockström et al., 2009a; Steffen et al., 2015). Political and societal developments during the 21st century and their feedback interactions with the planetary climate and biophysical environment will be decisive for the future trajectory of the Earth system (Lenton and Latour, 2018; Steffen et al., 2018). Business-as-usual is taking the planet into a 'hothouse Earth' state unprecedented for millions of years in geological history (Winkelmann et al., 2015; Ganopolski et al., 2016), while calls for rapid decarbonisation of the global economic system to meet the Paris climate agreement (Rockström et al., 2017) will also have complex consequences involving an intensified entanglement of social, economic and biophysical processes and their resulting feedback dynamics, up to the planetary scale (Mengel et al., 2018). Despite extensive debate about the Anthropocene (Lewis and Maslin, 2015; Hamilton, 2015; Brondizio et al., 2016; Zalasiewicz et al., 2017), and growing recognition of the limitations of current Earth system models for analysis and policy advice in the context of these shifting dynamics (van Vuuren et al., 2012, 2016; Verburg et al., 2016; Donges et al., 2017a, b; Calvin and Bond-Lamberty, 2018), little has been done to address the fundamental challenge of systematically reviewing the conceptual foundations of Earth system modelling to include dynamic social processes, rather than externalising them (Bretherton et al., 1986, 1988).

To understand planetary-scale social-ecological dynamics, models of World-Earth systems are urgently needed (Schellnhuber, 1998, 1999; Rounsevell et al., 2014; van Vuuren et al., 2016; Verburg et al., 2016; Donges et al., 2017a, b, 2020; Calvin and Bond-Lamberty, 2018). Epistemologically, we conceptualise World-Earth systems as planetary-scale systems consisting of the interacting biophysical subsystems of the Earth, and the social, cultural, economic, and technological subsystems of the World of human societies. It should be noted here that in the context of global change analysis and modelling, the term 'Earth system' was intended to include human societies and their activities and artefacts (Bretherton et al., 1988; Schellnhuber, 1998, 1999). However, in currently influential science and policy contexts, notably the Intergovernmental Panel on Climate Change (IPCC) (Flato, 2011; Flato et al., 2013), 'Earth system models' deal only with the physical dynamics of the atmosphere, ocean, land surface and cryosphere, and a limited set of interactions with the biosphere. While some might see tautology in the term

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).



‘World-Earth systems’, we use it to highlight that human societies, their cultures, knowledge and artefacts (the ‘World’) should now be included on equal terms in a new family of models to conduct systematic global analyses of the Anthropocene. A fully co-evolutionary approach is needed, in the sense of representing social-ecological feedback dynamics across scales.

Future World-Earth modelling efforts will largely be pieced together from existing conceptualisations and modelling tools and traditions of social and biophysical subsystems, which encode the state of the art in our understanding of the Anthropocene. Current efforts in World-Earth systems modelling are highly stylised (e.g. Kellie-Smith and Cox (2011); Garrett (2015); Jarvis et al. (2015); Heck et al. (2016); Nitzbon et al. (2017); Strnad et al. (2019)), or tend to be proof-of-concept prototypes (Beckage et al., 2018; Donges et al., 2020). None operate yet in a process-detailed, well-validated and data-driven mode. To serve these nascent efforts in enabling World-Earth systems analysis of the Anthropocene, this article addresses the core question of which are the relevant categories within which World-Earth models, as essential scientific macroscopes (Schellnhuber, 1999), should operate. The problem for both scientific integration and real-world application is that the characteristic basis of the interactions of social and biophysical subsystems is often not explicit in current models. Often, the interactions between these subsystems are not recognised at all. By framing a taxonomy around the current dominant distinctions – and disciplinary divides – we can begin to explore links and feedback mechanisms between taxa in more structured, systematic and transdisciplinary ways. With this taxonomy, we develop initial tools and terminologies that enable model builders and model users to be clear about their social, cultural, epistemological and perhaps also axiological standpoints.

We want to emphasise that this taxonomic approach does not presuppose that there is “one world” (an ontological position) when models of different worlds are combined, nor do we intend it to serve as a universal blueprint for models of essentially everything. Instead, we argue that a taxonomy can help to focus modellers’ attention better on the ontological and epistemic commitments within their models. This approach opens Earth system analysis to deeper dialogues with proponents of non-human actors as shapers of the world (Latour, 2017; Morton, 2013), or even the possibility of no world at all (Gabriel, 2013).

While the present article proposes a conceptual basis for World-Earth modelling, the proposed taxonomy is employed in the companion paper by Donges et al. (2020) to develop the operational World-Earth modelling framework copan:CORE. Here, this framework is cast into software and applied to construct and study an example of a novel World-Earth model that seeks to overcome the long-standing challenge of endogenising the choice of discount factors (describing how much societies value the present relative to the future) in climate mitigation studies.

1.2 Structuring the landscape of global environmental change models

Diverse scientific modelling communities aim to capture different aspects of social-ecological dynamics embedded in the Earth system up to planetary scales. Some processes operating in the Earth system are commonly described as being governed by the "natural laws" and generalizable principles of physics, chemistry and (to some extent at least) ecology (for example, atmosphere and ocean circulation as governed by the physical laws of fluid and thermodynamics), while others are thought to be dominated by human behaviour, decision making and collective social dynamics (e.g., the regularities underlying individual and social learning). This tendency for separate treatment of these different kinds of process in the natural and social sciences gives rise to problems when dealing with the many real-world subsystems that operate in both domains simultaneously. What

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

is more, different scientific communities use different methods and adhere to different viewpoints as to the nature and character of such subsystems and their interactions. There is now a number of conceptualisations of social-ecological or coupled human-environment systems in environmental, sustainability and Earth system science (e.g. Vernadsky (1929/1986); Schellnhuber (1998); Fischer-Kowalski and Erb (2006); Jentoft et al. (2007); Biggs et al. (2012)) but we see a pressing need to structure modelling efforts across communities, providing a joint framework while maintaining the conceptual flexibility required for successful cross-disciplinary collaboration.

Here, we propose a taxonomic framework for structuring the multitude of subsystems that are represented in current mathematical and computer simulation models. The motivation for proposing such an ordering scheme is:

1. to provide the means for collecting and structuring information on what components of social-ecological systems relevant to global change challenges are already present in models in different disciplines,
2. to point out uncharted terrain in the Earth system modelling landscape, and
3. to provide the foundations for a systematic approach to constructing future co-evolutionary World-Earth models, where feedback mechanisms between components can be traced and studied. This conceptual work aims to contribute to a central quest of sustainability science (Mooney et al., 2013) that "seeks to understand the fundamental character of interactions between nature and society." (Kates et al., 2001).

1.3 Definitions and explanations of key terms

In this article, we use the term subsystem to refer to any dynamic component in models of World-Earth systems. In this broad category, we can include both the kinds of subsystems that are governed mainly by "natural laws" of physics, chemistry or ecology (e.g., seasonal precipitation, ocean nutrient upwelling) and those that are governed mainly by human behaviour, decision making and collective social dynamics (e.g., international food trade, carbon taxes). Many scientific communities similarly make this distinction between biophysical ("natural", ecological, environmental) subsystems and socio-cultural (social, human, "anthroposphere") subsystems. We also highlight socio-metabolic subsystems at the overlap of societal and natural "spheres" of the Earth system (Fig. 1). We suggest that explicit attention to these subsystems and their interactions is needed in order to deepen the understanding of transformative change in the planetary social-ecological system, making a valuable contribution to the design and operational development of future, more comprehensive World-Earth models for charting sustainability transitions into a safe and just operating space for humanity (Rockström et al., 2009a; Raworth, 2012; Dearing et al., 2014).

A further note on the term *biophysical*: here, we use this word as a shorthand term to refer to Earth's interacting living and non-living components, encompassing geophysical (climatic, tectonic, etc.), biogeophysical, biogeochemical and ecological processes. These categories are significant in Earth system science because feedbacks involving these processes tend to have different dynamic characteristics. Accordingly, they have been dealt with very differently in Earth system analysis and modelling (Charney et al., 1977; Gregory et al., 2009; Stocker et al., 2013).

The co-evolution of Earth's geosphere and biosphere is a central concept in Earth system science (Lovelock and Margulis, 1974; Budyko et al., 1987; Lovelock, 1989; Schneider et al., 2004; Lenton et al., 2004; Watson, 2008), but the global models that currently dominate the field represent just a snapshot of the system, focused on the biophysical dynamics that play out over decades to centuries. We use the term co-evolution to describe the complex dynamics that arise from the reciprocal interactions of subsystems, each of which changes the conditions for the future time evolution of the other (not excluding, but not limited to processes of Darwinian co-evolution involving natural selection). Earth system models (ESMs) include key physical feedbacks, and increasingly permit the investigation of biophysical feedbacks, but as we have indicated, they lack socio-metabolic and socio-cultural subsystems, relying on narrative-based inputs for dealing with anthropogenic changes. Integrated assessment models (IAMs) used in the global change context (Edenhofer et al., 2014; van Vuuren et al., 2016) include some interactions of social and biophysical subsystems in order, say, to assess potential economic consequences of climate change and alternative climate policy responses. But they lack the kinds of interactions and feedback mechanisms (e.g., by impacts of climatic changes on socio-metabolic subsystems, or by the effects of socio-cultural formation of public opinion and coalitions in political negotiations on environmental policies) that societies throughout history have shown to be important which is revealed, e.g., by studies of social-ecological collapse and its connection to past climate changes (Weiss and Bradley, 2001; Ostrom, 2009; Donges et al., 2015; Cumming and Peterson, 2017; Barfuss et al., 2020). To explore and illustrate the consequences of these typically neglected interactions and feedbacks, we have studied a conceptual model that gives rise to complex co-evolutionary dynamics and bifurcations between qualitatively different system dynamics: a model of socially transmitted discount rates in a greenhouse gas emissions game, discussed in Section 4.

For completeness, we also provide brief definitions of our working terminology: a “link” or “interaction” is a causal influence of one subsystem on another that is operationally non-decomposable into smaller links; a “mechanism” is a micro-description of how exactly this causal influence is exerted; a “process” is a set of links that “belong together” from some suitable theoretical point of view; a “loop” is a closed path in the network of links; and an “impact” of a link is the change in the target system attributable to this link.

We should note here that this taxonomy is dealing with causal narratives from different scientific disciplines that are encoded in models, and as such, it does not require any a priori theories and hypotheses about causality. Causal narratives are our starting point because they are necessary for and are explicitly encoded in simulation modelling - and our classification lets us interrogate them more systematically and exposes them explicitly.

2 A taxonomy of subsystems in World-Earth systems models

In this section, we introduce the biophysical (ENV), socio-metabolic (MET), and socio-cultural (CUL) taxa for classifying subsystems in models of World-Earth systems (Fig. 1). For each taxon, we give examples of corresponding subsystems from different modelling fields. We also discuss how the suggested taxonomy relates to earlier conceptualisations of human societies embedded in and interacting with environmental systems (Sect. 2.4).

We have followed three guidelines in constructing this taxonomy for models of World-Earth systems:

1. *Compactness*, because we aim at a “top-level" framework that is useful and tangible, with as few classifications as possible, covering the scope of co-evolutionary modelling research parsimoniously and in a self-containing way.

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

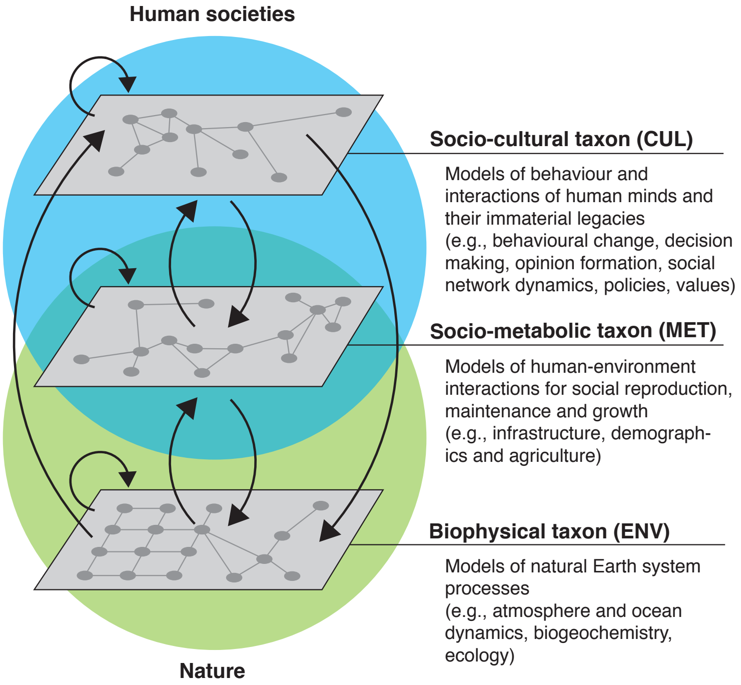

Human societies
Socio-cultural taxon (CUL)
Models of behaviour and interactions of human minds and their immaterial legacies (e.g., behavioural change, decision making, opinion formation, social network dynamics, policies, values)
Socio-metabolic taxon (MET)
Models of human-environment interactions for social reproduction, maintenance and growth (e.g., infrastructure, demographics and agriculture)
Biophysical taxon (ENV)
Models of natural Earth system processes (e.g., atmosphere and ocean dynamics, biogeochemistry, ecology)
Nature

Figure 1. Proposed taxonomy of subsystems in World-Earth systems models. The blue and green overlapping discs represent the current discipline-based domains in which the subsystems and processes of nature, human societies, and their interactions are modelled. Our scheme structures this continuum into three taxa (light grey layers) for model subsystems (dark grey discs): (i) a biophysical taxon (ENV), (ii) a socio-metabolic taxon (MET), and a socio-cultural taxon (CUL). Links within and between these modelled subsystems (shown as black arrows in the figure) can further be classified using a 3 × 3 taxonomy of interactions (Fig. 2, Sect. 3).

2. *Compatibility* with existing disciplines and research fields within, between and beyond the persistent natural/social sciences divide, because we view the scientific endeavour of understanding links and feedbacks in co-evolutionary World-Earth systems as an integrative and transdisciplinary opportunity.

3. *Operative capacity* for model classification and construction, because we want to advance efforts rapidly in World-Earth modelling. This guideline differs from the previous two in that it deals with practical aspects of modeling. We include it because it flags the need for critical reflection on the suitability of combined models for the tasks at hand. We want to be able to expand the scope of modelling to be more inclusive, allowing more differentiation and well-founded permutations of approaches.

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

Models encode knowledge outside of the mind of the modeller, so these guiding principles are intended to ensure that bridging across currently very distinct modelling approaches still permits tracing back how the techniques relate to the theories, assumptions, and framings of the contributory disciplines.

The proposed taxonomy reflects the longstanding structure – and the underlying divides – of the scientific disciplines dealing with the respective subsystems. We argue that it also provides a blueprint for navigating the fragmented modelling landscape and bringing new opportunities for cross-disciplinary bridging. The anthropocentric and dialectic distinction between the realms of nature or "the environment" and of human societies has a long intellectual history. Deep philosophical and scientific puzzles are connected with the attempts to draw a sharp distinction between these domains, and to satisfactorily integrate properties such as mental states, intentions, and life itself.

With the progressive improvements in biophysical Earth system modelling (Reichler and Kim, 2008; Steffen et al., 2020) and the concomitantly growing reliance on model-based insights for global decision-making over a wider range of urgent sustainability issue (National Research Council, 2007; Rounsevell et al., 2014; Calder et al., 2018), as is the case for example for the Paris climate agreement (UNFCCC, 2015) informed by the IPCC (Stocker et al., 2013; Barros et al., 2014; Edenhofer et al., 2014) and the policy processes derived from it, these conceptually challenging issues can now have direct practical implications. Illustration such different conceptions of Earth system processes, in models of the contemporary Earth system, land vegetation can be treated as inanimate carbon, a transpiration "pump" affecting precipitation and soil moisture patterns (e.g. Sitch et al. (2003)), or as the animate matter of biodiverse ecosystems that sustain human communities (e.g. (Purves et al., 2013)). Similarly, different assumptions in models about non-material factors such as human rationality, cognition, motivations, institutions and social connections lead to very different likelihoods for alternative sustainability pathways for the world's economies and material resource use (Donges et al., 2017b; Müller-Hansen et al., 2017; Beckage et al., 2018; Otto et al., 2020b).

For these reasons, we follow a pragmatic approach in proposing a taxonomic framework that draws upon examples and allows for overlap between the domains of nature and human societies, where materiality meets intention (noting that in complex social-ecological systems, purposeful intervention will be accompanied by unintended or unanticipated side effects). Following this approach, modelled subsystems in the biophysical taxon are situated in the material domain of nature, those in the socio-metabolic taxon lie in the overlap domain, and those in the socio-cultural taxon reside in the immaterial domain of human cultures (Fig. 1).

2.1 Biophysical taxon

The biophysical taxon (ENV) contains the processes and subsystems that are typically included in current comprehensive Earth system models, but views them from the perspective of the Anthropocene shift to human "co-control". These subsystem models are governed by deterministic and stochastic mathematical equations, often developed from first principles about the physical relationships involved. There is a case for subdividing the biophysical taxon into an ecological subtaxon (subsystems associated with life) and a geophysical subtaxon (subsystems not associated with life), since they have distinct, albeit co-evolving dynamics (Vernadsky, 1929/1986; Lenton et al., 2004), and this subdivision would correspond to widely accepted

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

geosphere/biosphere conceptualisations of the Earth system (Bretherton et al., 1986, 1988; Seitzinger et al., 2015). However, we apply our principle of compactness, because geosphere-biosphere links and processes have been comprehensively documented over the past few decades, as they underpin current Earth system and global integrated assessment modelling. Rather than retracing these links (after all, the existing models are not going to be completely reconfigured in light of the issues we explore in this paper), we have opted to take today's state of the art in biophysical global modelling as our main point of departure, following the principle of compatibility introduced above.

Earth system models have developed from coupled atmosphere-ocean general circulation models, progressively coupling in components describing biogeochemical and biogeophysical dynamics. On decade-to-millennium time scales relevant for the analysis of anthropogenic climate change and its medium-term consequences, examples of these modelled subsystems where human-controlled dynamics are prominent concerns include atmospheric chemistry, ocean productivity, sea ice, land vegetation, and major elemental cycles such as those of nitrogen, phosphorus, and sulfur (Bretherton et al., 1986, 1988). Furthermore, as it becomes clearer that palaeoclimate models can play a vital role in "deep future" studies of human-controlled processes in the Anthropocene, Earth system dynamics operating on longer time-scales are relevant (Zeebe and Zachos, 2013; Steffen et al., 2018). So for these purposes, the biophysical taxon would include subsystems involving the lithosphere (e.g., rock weathering, isostatic depression and rebound associated with the advance and retreat of ice sheets on land) and even external drivers such as large-body impacts (Brugger et al., 2017), if these provide "natural experiments" or analogues for future change.

Research fields dealing with models of subsystems belonging to the biophysical taxon include, among others, geophysics, meteorology, oceanography, biology, ecology, biogeochemistry, and geology. Few of these sciences have yet grasped the methodological and theoretical tools for dealing with the human dimensions of anthropogenic change. From our planetary-scale perspective, the ENV taxon exhibits a substantial overlap with categories such as models of "the environment", "nature" or "ecology", with their specific disciplinary connotations, although many of these models have tended to be small-scale, context-specific and idiographic. An exception from this are global dynamic vegetation models such as LPJ (Sitch et al., 2003), which focus, however, on representing the physical dynamics of ecological processes and structures in an Earth system context and not on ecological dynamics as such (i.e., interactions between living organisms). We note a current drive for further refinements of ecological dynamic network processes in large-scale modelling (Purves et al., 2013; Harfoot et al., 2014) within the ENV taxon that may improve global-scale conceptualisations of ecosystems in ways compatible with both Earth system modelling and socio-ecological systems research and resilience thinking.

2.2 Socio-metabolic taxon

The socio-metabolic taxon contains processes and subsystems that form the material basis and products of societies, making direct interconnections between human societies and the biophysical environment that sustains them. This taxon comprises models of demographics and social structure (e.g., population size, age/sex distribution, health parameters; and social categories with material or resource-use consequences, such as class, clan, caste, ethnicity). It also includes "the technosphere": society's artefacts, factors of production and technologies (e.g. labour, land, capital, natural resources, raw material, energy;

tools, machines, infrastructure; cultivated landscapes, domesticated animals and plants), and economic systems (manufacturing, distribution and consumption of goods and services) (Haff, 2012, 2014; Mooney et al., 2013; Herrmann-Pillath, 2018).

The broad field of economics currently dominates descriptions of parts of the socio-metabolic taxon in quantitative models, but many other disciplines such as geography, industrial metabolism, social ecology, and science and technology studies also play a role. In modelling terms, this taxon typically involves representations of both the biophysical planet Earth and the socio-cultural World of human societies. This implies hybrid models of the type that are currently included in Integrated Assessment Models of global change, and entails strong simplifying assumptions. We suggest that our approach can bring much-needed clarity and transparency about the role of such models in understanding World-Earth systems (c.f. van Vuuren et al. (2016)). One should note that IAMs and economic models are typically expressed in terms of financial value and not material flows that directly interact with subsystems in ENV (mostly empirical input-output theories of economics being an exception, Leontief (1936)).

2.3 Socio-cultural taxon

The socio-cultural taxon contains processes and subsystems that are described in models of the behaviour of human minds and their immaterial legacies, abstracted from their biophysical foundations and often described as lying in the realm of human agency (Otto et al., 2020b). Of the three taxa proposed, processes and subsystems in the socio-cultural taxon are the least formalised in mathematical and computer simulation models so far, despite substantial efforts in this direction in many fields of the social sciences (e.g. Farmer and Foley (2009)) and a likelihood that they may be only partly formalizable. Research fields dealing with models of processes and subsystems in the socio-cultural taxon include sociology, anthropology, behavioural economics, political science and social ecology. Our taxonomic approach can enable the diverse modelling activities now underway to engage more directly with the incipient World-Earth modelling effort.

Examples of modelled subsystems in this taxon include individual and collective opinions, behaviours, preferences, and expectations and their social network dynamics; information and communication networks; institutions and organisations; financial markets and trade; political processes; social norms and value systems (Mooney et al., 2013). Notably, the CUL taxon can also include processes of digital transformation and artificial intelligence that increasingly restructure and shape the socio-cultural sphere of human societies. It also provides a locus for debating the challenge of reflexiveness in science, especially in fields where modelling plays a vital role in shaping knowledge and action (Yearworth and Cornell, 2016). For instance, future World-Earth modelling will have to grapple with ways to recognize Earth system science as an endogenous generator of scientific conceptions of 'Earth'. Relevant for modelling efforts, socio-cultural subsystems can vary on substantially different time scales. Near instantaneous information exchanges are possible on online social networks and within and between increasingly advanced algorithms (e.g. algorithmic trading systems on financial markets), while elections and governance processes act on the order of years. Formal institutions (e.g. laws) change on the order of decades and informal institutions (e.g. religions) develop over time frames on the order of centuries to millennia (Williamson, 1998; Otto et al., 2020a).

2.4 Relations to other conceptualisations of social-ecological systems

Our model-centred taxonomy is inspired by previous systemic conceptualisations of human societies embedded in the Earth system, building upon them in a way that may help to bridge across diverse disciplines and theoretic traditions.

In one of the earliest Earth system conceptualisations, Vernadsky (1929/1986) distinguishes the inanimate matter of the geosphere, the living biosphere, and the noosphere of networked consciousness, the latter reverberating in recent conceptualisations of the technosphere and planetary human-Earth system interactions (Herrmann-Pillath, 2018; Lenton and Latour, 2018). Along these lines, Schellnhuber (1998, Fig. 34) introduced the ecosphere (directly corresponding to our ENV taxon, entailing geophysical and ecological interactions), the anthroposphere (broadly related to MET, but with some socio-cultural features), and the global subject (closely related to CUL).

Conceptualisations in resilience theory, ecological economics and sustainability science emphasise the interactions and interdependence of biosphere and society (Brundtland, 1987; Folke, 2006; Folke et al., 2011), with many sustainability practitioners adding the economy to make “three pillars” or a “pie of sustainability" consisting of economy embedded in society embedded in biosphere (Folke et al., 2016). These fields have typically focused on local to regional geographic scales or specific sectors, and have not placed much emphasis on global modelling, but in general terms, their view of society contains aspects of our MET taxon, while “the economy” is more restricted than MET. Herrmann-Pillath (2020) argues that the field of ecological economics would benefit from more attention to the creative processes of ‘art’, which we would frame as CUL aspects that are largely absent from current conceptualisations in that field and also more broadly (as also argued by Jax et al. (2013); Woroniecki et al. (2020)).

Fischer-Kowalski and Erb (2006) explicitly develop the concept of social metabolism, in terms of the set of flows between nature and culture, in order to describe deliberate global sustainability transitions. Governance-centred classification schemes in social-ecological systems research (Jentoft et al., 2007; Biggs et al., 2012), in the tradition of Ostrom (Ostrom, 2009), can also be brought into our taxonomy. Categories of the governance (sub)system link CUL and MET, and the (sub)system to be governed (ENV and MET) links the biophysical resources to be used with the social agents who will use them.

The taxonomy approach means that things that were previously included in models as opaque and unquestioned systems can be unpacked and critically examined. This would be of particular benefit to model users who were not the model builders. For example, education may be explicitly linked to demography (as in various integrated assessment models), so typically would be treated as a quantifiable and accumulable process in the MET taxon: i.e., investment in women’s education results in a lower birth rate and therefore less future land use. In CUL, education would perhaps be treated in a more relational way - dealing with the spread of ideas, development of communities, changes in power structures etc.

3 Taxonomy of subsystem interactions in World-Earth systems models

In this section, we describe a taxonomy of modelled interactions between subsystems that builds upon the taxonomy of subsystems. The three taxonomic classes for World-Earth subsystems give rise to nine taxa for directed interactions connecting these subsystems. Given a pair of taxonomic classes of subsystems A and B, the taxonomic class for directed interactions

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

between A and B is denoted as $A \rightarrow B$. Here, a directed interaction is understood in the sense of a modelled subsystem in A exerting a causal influence on another modelled subsystem in B. For example, greenhouse gas emissions produced by an industrial subsystem in MET that exert an influence on the Earth's radiative budget in ENV would belong to the interaction taxon MET → ENV. Three of the nine interaction taxa correspond to self-interactions within taxa, while six interaction taxa connect distinct subsystem taxa (Fig. 2).

In the following, we focus on describing examples of such modelled interactions between pairs of subsystems that are potentially relevant for future trajectories of World-Earth systems in the Anthropocene and give examples of published models containing them. The content presented in the subsections necessarily differs in scope and depth reflecting today's dominant modelling priorities, but we have aimed to ensure the information is comparable. All subsections below provide (i) a general description of the interaction taxa with some examples, and (ii) a summary of how these interactions are represented in current models.

Furthermore, possible extensions of our taxonomic approach to classify feedback loops and more complex interaction networks between subsystems are discussed (Sect. 3.10). We acknowledge that finding a conceptualisation that is satisfactory for all purposes is unlikely, but our particular pragmatic taxonomy can be useful for constructing models of World-Earth systems. It has already proven fruitful in the development of the copan:CORE open World-Earth modelling framework (Donges et al., 2020) by guiding the choice of process classes and entities that can be described in the framework as well by defining the coupling interfaces of model components that can be integrated using copan:CORE.

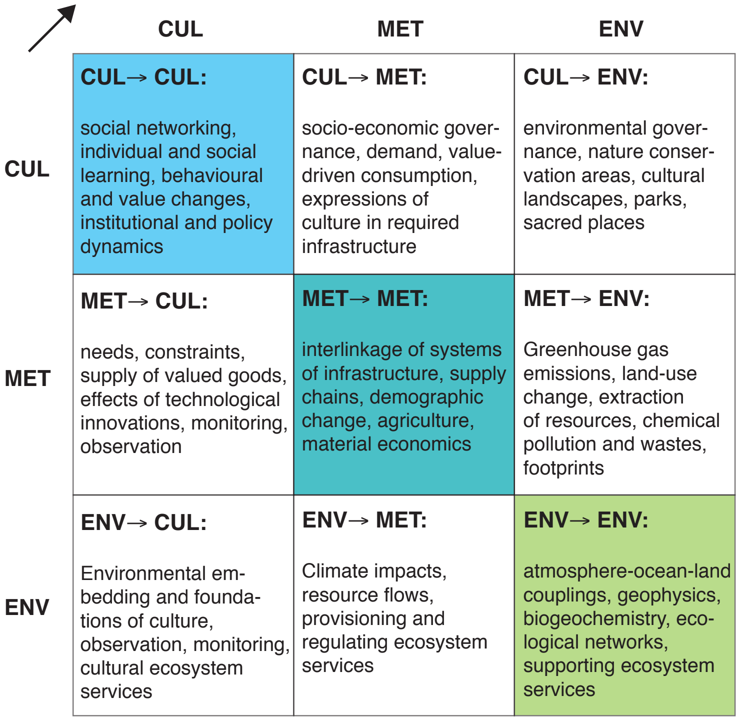
3.1 ENV → ENV: Biophysical Earth system self-interactions

This taxon encompasses interactions between biophysical subsystems of the type studied in current process-detailed Earth system models such as those in the CMIP5 model ensemble (Taylor et al., 2012) used in the IPCC reports (Stocker et al., 2013). For example, this includes modelled geophysical fluxes of energy and momentum between atmosphere and ocean, interactions between land vegetation, atmospheric dynamics and the hydrological cycle, or, more generally, exchanges of organic compounds between different compartments of biogeochemical cycles (excluding human activities here).

A detailed representation of these biophysical interactions is largely missing so far in current first attempts at modelling social-ecological dynamics at the planetary scale (e.g. Kellie-Smith and Cox (2011); Heck et al. (2016)). However, emerging socio-hydrological (Di Baldassarre et al., 2017; Keys and Wang-Erlandsson, 2017) and agent-based land-use dynamics models at regional scales (Arneth et al., 2014; Rounsevell et al., 2014; Robinson et al., 2017) include some processes involving interactions between biophysical subsystems such as the atmosphere, hydrological cycles and land vegetation.

3.2 ENV → MET: Climate impacts, provisioning and regulating ecosystem services, etc.

This taxon describes modelled interactions through which biophysical subsystems exert an influence on socio-metabolic subsystems. Relevant examples in the context of global change in the Anthropocene include the impacts of climate change on human societies (Barros et al., 2014) such as damages to settlements, production sites and infrastructures and supply chains (Otto



↗	CUL	MET	ENV
CUL	**CUL→ CUL:** social networking, individual and social learning, behavioural and value changes, institutional and policy dynamics	**CUL→ MET:** socio-economic governance, demand, value-driven consumption, expressions of culture in required infrastructure	**CUL→ ENV:** environmental governance, nature conservation areas, cultural landscapes, parks, sacred places
MET	**MET→ CUL:** needs, constraints, supply of valued goods, effects of technological innovations, monitoring, observation	**MET→ MET:** interlinkage of systems of infrastructure, supply chains, demographic change, agriculture, material economics	**MET→ ENV:** Greenhouse gas emissions, land-use change, extraction of resources, chemical pollution and wastes, footprints
ENV	**ENV→ CUL:** Environmental embedding and foundations of culture, observation, monitoring, cultural ecosystem services	**ENV→ MET:** Climate impacts, resource flows, provisioning and regulating ecosystem services	**ENV→ ENV:** atmosphere-ocean-land couplings, geophysics, biogeochemistry, ecological networks, supporting ecosystem services

Figure 2. Taxonomic matrix for classifying directed interactions between subsystems in World-Earth systems models. This 3×3 classification system builds upon the taxonomy of three classes for subsystems introduced in Sect. 2. The unshaded matrix elements (here containing examples of interactions) correspond to the interaction arrows drawn between the three subsystem taxa shown in Fig. 1. Shaded elements correspond to self-interactions. The examples for directed interaction mechanisms given in the matrix elements are indicative and based on our particular areas of research.

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

et al., 2017), impacts on agriculture or human health, but also provisioning and regulating ecosystem services such as resource flows (Millennium Ecosystem Assessment, 2005).

Some of these interactions such as climate change impacts are now being included in IAMs (a prominent example being the DICE model, Nordhaus (1992)) and stylised models (for example Kellie-Smith and Cox (2011) and Sect. 4), but there remain challenges, e.g. in estimating damage functions and the social cost of carbon (Nordhaus, 2017). Influence from weather and climate on agriculture are studied on a global scale using model chains involving terrestrial vegetation models such as LPJ (Sitch et al., 2003) and agricultural economics models such as MAgPIE (Nelson et al., 2014). As another example, models of the distribution of vector-born diseases such as Malaria are employed to assess the impacts of climate change on human health (Caminade et al., 2014).

3.3 ENV → CUL: observation, monitoring, cultural ecosystem services, etc.

This taxon contains modelled interactions through which the state of the biophysical environment directly influences socio-cultural subsystems. These links can be mediated through the observation, monitoring and assessment of environmental change from local to global scales (e.g., chemical pollution, deforestation or rising greenhouse gas concentrations in the atmosphere) by social actors that in turn are processed by public opinion formation and policy making in socio-cultural subsystems (Mooney et al., 2013). The links described by the ENV → CUL taxon also relate to cultural identity connected to the environment, sense of place (Masterson et al., 2017), and more generally what has been described as cultural ecosystem services (Millennium Ecosystem Assessment, 2005). For example, Beckage et al. (2018) have modelled the effect of changes in extreme events resulting from climate change on risk perception of individuals. Changes in risk perception may result in changes in emission behaviour given the perceived behaviour of others (social norms) and structural conditions in society, thus feeding back on future climate change.

ENV → CUL also play a role in regional-scale models of poverty traps where decline in natural capital reduces traditional ecological knowledge as a form of cultural capital (Lade et al., 2017b), or in models of human perceptions of local scenic beauty in policy contexts (Bienabe and Hearne, 2006). At the moment, most models deal with these interactions only at a sub-global level. But there is increasing recognition of the need for the more dynamic understanding that formal modelling can provide of such complex psychologically and culturally mediated aspects of human behavior in the Anthropocene (Schill et al., 2019).

3.4 MET → MET: economic and socio-metabolic self-interactions

This taxon describes modelled interactions between MET subsystems that connect the material manifestations and artefacts of human societies. Examples include the energy system driving factories, supply chains connecting resource extractors to complex networked production sites or machines constructing infrastructures such as power grids, airports and roads.

Certain processes involving such interactions, e.g. links between the energy system and other sectors such as industrial production, are represented in IAMs in an abstracted, macroeconomic fashion. There exist also agent-based models resolving the dynamics of supply chains that allow to describe the impacts of climate shocks on the global economy in much more detail (e.g. Otto et al. (2017)). Another class of examples are population models that may include factors such as the influence of income on fertility (Lutz and Skirbekk, 2008). However, to our best knowledge, process-detailed models of the socio-industrial metabolism (Fischer-Kowalski and Hüttler, 1998; Fischer-Kowalski, 2003) or the technosphere (Haff, 2012, 2014) comparable in complexity to biophysical Earth system models have not been published so far.

3.5 MET → ENV: greenhouse gas emissions, land-use change and biodiversity loss, impacts on other planetary boundary processes, etc.

This taxon encompasses modelled influences exerted by socio-metabolic subsystems on the biophysical environment including various forms of the "colonisation of nature" (Fischer-Kowalski and Haberl, 1993). Prominent examples in the context of global change and sustainability transformation include human impacts on the environment addressed by the planetary boundaries framework (Rockström et al., 2009a, b; Steffen et al., 2015) such as anthropogenic emissions of greenhouse gases (Stocker et al., 2013), nitrogen and phosphorous, other forms of chemical pollution and novel entities (e.g., nano particles, genetically engineered organisms), land-use change and induced biodiversity loss, exploitation and use of natural resources (Perman, 2003). This taxon also includes various forms of the conversion of energy and entropy fluxes in the biophysical Earth system by human technologies such as harvesting of renewable energy by wind turbines and photovoltaic cells (Kleidon, 2016) or different approaches to geoengineering (Vaughan and Lenton, 2011).

The interactions described by the MET → ENV are central in IAM and ESM studies of the global environmental impacts of human activities in the Anthropocene such as anthropogenic climate change as driven by greenhouse gas emissions and land-use change (Barros et al., 2014; Edenhofer et al., 2014). The latter two key processes are also frequently included in emerging studies of planetary social-ecological dynamics using stylised models (Kellie-Smith and Cox, 2011; Anderies et al., 2013; Heck et al., 2016; Heitzig et al., 2016; Lade et al., 2017a; Nitzbon et al., 2017).

3.6 MET → CUL: needs, constraints, etc.

This taxon describes modelled influences and constraints imposed upon socio-cultural dynamics by the material basis of human societies (socio-metabolic subsystems). These include, for example, the effects, needs and constraints induced by the biophysical "hardware" that runs socio-cultural processes: infrastructures, machines, computers, human bodies and brains, and associated availability of energy and other resources. It also includes the effects of technological evolution, revenues generated from economic activity, supply of valued goods, e.g. on opinion formation and behavioural change in the socio-cultural domain, or the consequences of change in demographic distribution of pressure groups on political systems and institutions.

As a recent example, the Beckage et al. (2018) model mentioned above (Sect. 3.3) has one parameter to reflect structural constraints in society that affects the degree to which emission behaviour can be changed. MET → CUL links also appear in models of resource use in social-ecological systems, where social learning of harvesting effort depends on the harvest rate (Wiedermann et al., 2015; Barfuss et al., 2017; Geier et al., 2019) and fish catches influence perceptions about the state of the fishery (Martin and Schlüter, 2015; Lade et al., 2015), or in models of economic impacts on individual voting behaviour (Lewis-Beck and Ratto, 2013).

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

3.7 CUL → CUL: socio-cultural self-interactions

This taxon contains modelled self-interactions between subsystems in the socio-cultural domain that have been described as parts of the noosphere (Vernadsky, 1929/1986), the global subject (Schellnhuber, 1998), or the mental component of the Earth system (Lucht and Pachauri, 2004). Examples include the interaction of processes of opinion dynamics and preference formation on social networks, governance systems and underlying value systems (Gerten et al., 2018) as well as interactions between different institutional layers such as governance systems, formal and informal institutions (Williamson, 1998; Otto et al., 2020a).

Some of these processes related to human behaviour and decision making (Müller-Hansen et al., 2017) have already been studied in models of social-ecological systems on local and regional scales (Schlueter et al., 2012; Schlüter et al., 2017) and have been modelled in various fields ranging from social simulation to the physics of social dynamics (Castellano et al., 2009). However, they are so far largely not included in IAMs of global change or stylised models of planetary social-ecological systems (Verburg et al., 2016; Donges et al., 2017a, b).

3.8 CUL → ENV: environmental governance, nature conservation areas, social taboos, sacred places etc.

This taxon encompasses modelled influences that socio-cultural subsystems exert on the biophysical environment. An example for such a class of interactions is environmental governance realized through formal institutions (Ostrom et al., 2007; Folke et al., 2011), where a piece of land is declared as a nature protection area excluding certain forms of land-use which has a direct impact on environmental processes there. Similarly, nature protection areas for biodiversity conservation have been represented in marine reserve models (Gaines et al., 2010). Another related example for CUL → ENV links are nature-related values and informal institutions such as respecting sacred places in the landscape and following social taboos regarding resource use (Colding and Folke, 2001). Different forms of environmental governance have been modelled via so-called decision or sustainability paradigms (Schellnhuber, 1998; Barfuss et al., 2018; Heitzig et al., 2018).

Direct CUL → ENV links arguably cannot be found in the real world, in that socio-cultural influences on environmental processes must be mediated by their physical manifestations in the socio-metabolic domain (e.g. in the case of nature protection areas through the constrained actions of resource users, government enforcement efforts and infrastructures such as fences). However, such direct CUL → ENV links may be implemented in models, even on the global scale, such as in trade-off assessments of multiple land-uses (e.g. Boysen et al. (2017); Phalan (2018)).

3.9 CUL → MET: socio-economic policies and governance choices, value-driven consumption, etc.

Finally, this taxon contains modelled links pointing from socio-cultural to socio-metabolic subsystems. Examples include socio-economic policies and governance choices such as taxes, regulations or caps that influence the economy (e.g. carbon caps or taxes in the climate change mitigation context) or demographics (e.g. family planning and immigration policies) as well as the physical manifestations of financial market dynamics such as real estate bubbles. CUL → MET interactions also encompass the influence of cultural values, norms and lifestyles on economic demand and consumption and consequent changes in industrial production, building, transportation and other sectors.

Policy measures such as taxes, regulations or caps are much studied by IAMs of anthropogenic climate change (Edenhofer et al., 2014), while influences of value and norm change on economic activities such as general resource use (Wiedermann et al., 2015; Barfuss et al., 2017; Geier et al., 2019) and fishing (Martin and Schlüter, 2015; Lade et al., 2015) has been studied in the social-ecological modelling literature, but at a mostly local to regional level.

3.10 Higher-order taxonomies of feedback loops and more complex interaction networks

Beyond the taxonomy of interactions introduced above, higher-order taxonomies could also be derived. For example, a taxonomy of feedback loops can be derived from the 3×3 taxonomy of links, leading to six taxa for feedback loops of length two in models of World-Earth systems: given a pair of interaction taxa $A \to B$ and $B \to A$, the resulting taxon for loops between A and B may be denoted as $A \circlearrowleft B$. Many such feedback loops relevant for sustainability are not or only rigidly treated in current ESMs and IAMs. For example, the ENV $\circlearrowleft$ MET feedback loop is typically not sufficiently represented in IPCC-style analyses, because the impacts of climate change on human societies are not explicitly modelled or ill-constrained in IAMs (Sect. 3.5). Furthermore, feedback loops of the type CUL $\circlearrowleft$ X, where X may be subsystems from ENV, MET or CUL are mostly missing altogether, not the least because CUL is not represented, or only fragmentarily included, in current ESMs and IAMs.

Longer and more complex paths and subgraphs of causal interactions between subsystems could be classified by further higher-order taxonomies (e.g. inspired by the study of motifs, small subgraphs, in complex network theory, Milo et al. (2002)). This approach quickly leads to a combinatorial explosion, e.g. for 3-loops of the type $A \to B \to C \to A$ involving three modelled subsystems A, B, C and their interactions enumeration and counting of all possible combinations shows that there are already 11 distinct taxa for feedback loops of this kind. However, there are systematic methods available for classifying and clustering causal loop diagrams that could be leveraged to bring order into more complex models of World-Earth systems (Van Dijk and Breedveld, 1991; Rocha et al., 2015). Overall, such higher-order taxonomies could help in the design of models or model suites that can deal with different aspects of (nonlinear) interactions between World-Earth subsystems and serve as tools for understanding the emergent co-evolutionary macrodynamics.

4 An exemplary model showing complex co-evolutionary dynamics in a World-Earth system

At present, to our best knowledge, process-detailed World-Earth models that are comprehensive in the sense of the proposed taxonomies are not available. Therefore, in this section, we give an illustrative example of a stylised World-Earth system model that covers all classes of real-world processes that appear relevant in major global feedbacks. Even such a very simple World-Earth system model can contain a social-ecological feedback loop involving subsystem interactions introduced above (Sect. 3), and leading to a biophysical Earth system dynamics that depends crucially on a social-cultural evolution and vice versa. We also demonstrate how the taxonomies described above can be applied to classify model components and reveal the interaction

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

structures that are implicit in the model equations. The companion paper of this article applies the taxonomies to develop a more complex illustrative World-Earth model using the copan:CORE framework (Donges et al., 2020).

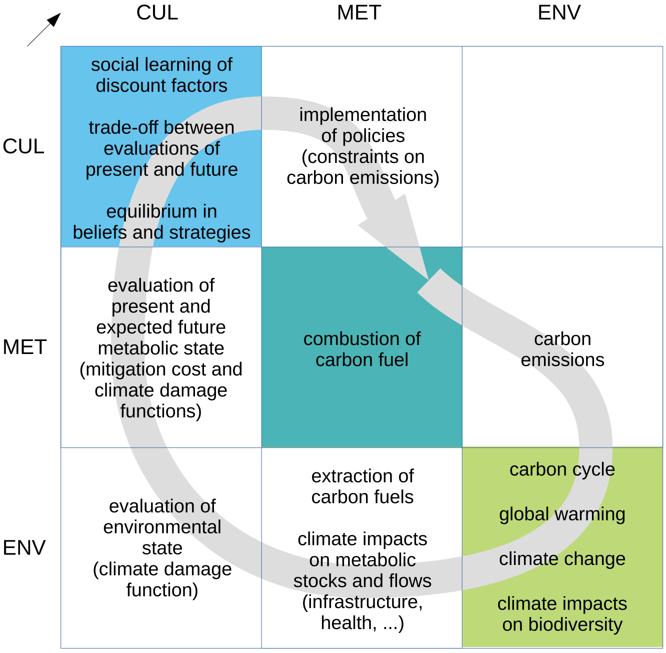
The example model studied here, copan:DISCOUNT, describes a world where climate change drives a change of countries' value systems, represented here just by the long-term discount factors their governments use in policy-making, which can be interpreted as their relative interest in future welfare as opposed to current welfare. These discount factors drive countries' emissions and thus in turn drive climate change, represented by a global atmospheric carbon stock. While the detailed description of the model's assumptions below will make clear that this causal loop involves eight of the nine interaction taxa shown in Fig. 2, the model is so designed that the description of the resulting dynamics from all these interactions can be reduced to just two ordinary differential equations, one for the fraction of "patient" countries and one for atmospheric carbon stock. The novelty of this model is that it endogenises socially transmitted choice of discount rates in a greenhouse gas emissions game to illustrate the effects of social-ecological feedback loops that are so far typically not considered in current climate economics and IAM modelling efforts.

The aim of this particular model design is to show clearly that while the taxonomy developed in this paper aims at being helpful in designing and analysing World-Earth models, this does not mean the different taxa need always be easily identifiable from the final model equations.

Before relating its ingredients to the introduced taxa, let us describe the model without referring to that classification. In our model, we assume that each country's metabolic activities are guided by a trade-off between the undesired future impacts of climate change caused by global carbon emissions, and the present costs of avoiding these emissions domestically. Similar to the literature on international environmental agreements and integrated assessment modelling, this tradeoff is modelled as a non-cooperative game between countries applying cost-benefit optimisation. The tradeoff and hence the evolution of the carbon stock is strongly influenced by the discount factor δ that measures the relative importance a country assigns to future welfare as compared to present welfare. The higher δ, the more a country cares about the future and the more they will reduce their emissions in order to avoid future climate impacts. While the economic literature treats δ as an exogenous parameter that has to be chosen by society (e.g., Arrow et al. (2013)), our model treats δ as a social trait that changes in individual countries over time because countries observe each other's welfare and value of δ and may learn what a useful δ is by imitating successful countries and adopting their value of δ. Because of the existence of climatic tipping points, this social dynamics does not only influence the state of the climate system but is in turn strongly influenced by it. Depending on whether the system is far from or close to tipping points, the trade-off between emissions reduction costs and additional climate damages can turn out quite differently and different values of δ will be successful.

Let us now present and decompose the model's basic causal loop in terms of the above introduced taxonomy, as shown in Fig. 3, starting in the central box. The countries' metabolisms (MET) combust carbon (MET $\rightarrow$ MET), leading to emissions (MET $\rightarrow$ ENV) that increase the global atmospheric carbon stock C (ENV), part of which is then taken up by other carbon reservoirs (ENV $\rightarrow$ ENV). C increases global mean temperature, leading to climate change (ENV $\rightarrow$ ENV) and thus to future climate impacts (i) on the countries' metabolisms (ENV $\rightarrow$ MET) and (ii) on aspects of the environment people care about, such as biodiversity (ENV $\rightarrow$ ENV $\rightarrow$ CUL). Countries evaluate these expected damages (MET $\rightarrow$ CUL; ENV $\rightarrow$ CUL)

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).



	CUL	MET	ENV
CUL	social learning of discount factors trade-off between evaluations of present and future equilibrium in beliefs and strategies	implementation of policies (constraints on carbon emissions)	
MET	evaluation of present and expected future metabolic state (mitigation cost and climate damage functions)	combustion of carbon fuel	carbon emissions
ENV	evaluation of environmental state (climate damage function)	extraction of carbon fuels climate impacts on metabolic stocks and flows (infrastructure, health, ...)	carbon cycle global warming climate change climate impacts on biodiversity

Figure 3. Planetary social-ecological processes and interactions represented in the copan:DISCOUNT model displayed in matrix form following Fig. 2. The co-evolutionary cycle of dynamic interdependencies implemented in the model is indicated by the grey arrow.

and the costs of avoiding emissions (MET → CUL), use their respective discount factors (CUL), which they learn by imitation (CUL → CUL), to assess possible domestic emissions constraints, then reach a strategic equilibrium with other countries (CUL → CUL) and implement the chosen emissions constraints (CUL → MET), this closing the long loop.

In the statistical limit of this model for a large number of countries, derived in detail in the Appendix A, this complex feedback dynamics is nicely reduced to just two equations,

$$\dot{C} = E_0 - c\,s(C)\,\phi(F) - rC, \tag{1}$$

$$\dot{F} = \ell F(1-F)[P(D(C,F)) - P(-D(C,F))], \tag{2}$$

where C is excess atmospheric carbon stock and F the fraction of "patient" countries (those that apply a large value of δ), and where $s(C)$ is a damage factor, $\phi(F)$ is a certain linear transformation of F, $D(C,F)$ is the utility difference between a country using discount factor α and a country using β, and $P(D)$ is a resulting imitation probability, all these derived in detail in the Appendix A. Some of the various terms in these formulas can be classified clearly as belonging to one taxon, e.g., BAU emissions E_0 belong to MET → ENV, carbon-uptake $-rC$ to ENV → ENV, and the imitation probability $P(D)$ to CUL →

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

CUL. But others cannot, e.g., certain terms occurring in the formula for D combine climate damages $s(C)$ (ENV → MET → CUL) with countries' values systems, represented by $\phi(F)$ (CUL). The dynamics are governed by about a dozen parameters controlling the relative speeds and intensities of subprocesses, costs and benefits of emissions reductions, and details of the learning-by-imitation process, as described in the Appendix (Sect. A).

Let us analyse a typical dynamics of the model, shown in Fig. 4, and relate it again to our taxonomy of subsystem interactions. Consider the middle green trajectories in the lower panel starting at a low atmospheric carbon stock of $C = 1$ (fictitious units) and a medium fraction of patient countries of $F = 0.5$ (green dot). At this point, both patient and impatient countries evaluate the state of the world very similarly, hence not much imitation of discount factors happens (weak CUL → CUL dynamics), so that F may fluctuate somewhat but is not expected to change much. At the same time, as the climate damage curve (middle panel) is still relatively flat, global emissions are higher than the natural uptake rate (strong MET → ENV influence), and C is likely to increase to about 1.7 without F changing much. During this initial pollution phase, climate damages increase (the ENV → MET/CUL links becomes stronger) and the slope of the damage curve increases as more climatic tipping points are neared or crossed. This decreases the patient countries' evaluations faster than the impatient countries', hence patience becomes less attractive and countries fatalistically decrease their discount factor, so that F declines to almost or even exactly zero (the CUL → CUL dynamics becoming first stronger then weaker again) while C grows to about 3.0. In that region, most tipping points are crossed and the damage curve flattens again, causing the opposite effect, i.e., making patience more attractive. If the idea of patience has not "died-out" at that point (i.e., F is still > 0), discount factors now swing to the other extreme with F approaching unity (CUL → CUL dynamics becoming temporarily very strong), shown by one green trajectory, while emissions are first almost in equilibrium with natural carbon uptake at about $C = 3.2$ (weak MET → ENV effect) and then decline ever faster once the vast majority of countries got patient (stronger MET → ENV). This trajectory finally converges to the stable steady state at a low carbon stock of about $C = 1.5$ and $F = 1$. Note that there is also some small probability that this point is reached much faster without the long detour if the stochastic social dynamics at the starting point give patience a random advantage, as on two of the plotted trajectories.

As is typical in models with various interactions, changes in their relative interaction rates can cause highly nonlinear and even qualitative changes in model behaviour. A comparison of the top and bottom panels in Fig. 4 (see also its caption) shows that this is in particular true for World-Earth models when the rates of socio-cultural processes of the CUL → CUL type are changed (as can be claimed is indeed happening in reality since the middle of the 20th century). It should be emphasised again that these socio-cultural processes are specifically those that are least or not at all represented in current models of global change, pointing to the necessity and expected progress in understanding when including them in more comprehensive World-Earth models.

Overall, the DISCOUNT model provides a first test of the taxonomy's guiding principles. It demonstrates the taxonomy's operative capacity to trace links between established dynamical systems methodology and macro behaviour; it is compatible with diverse research fields, here linking, among others, carbon cycles and social learning; and it has appropriate compactness, since tracing the loops and flows between taxa in this World-Earth model do not make us need to rethink the whole structure of the taxonomy.

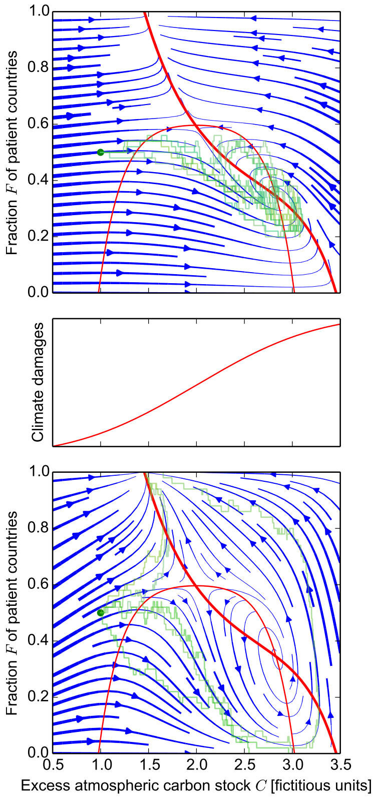


Figure 4. Typical dynamics of the copan:DISCOUNT model of the co-evolution of the global atmospheric carbon stock C and the time preferences of countries, represented by the fraction F of patient countries. Of five simulated stochastic trajectories (top and bottom panel, green lines) starting at the same initial state (green dot), some will converge fast to the more desirable stable steady state at $C \approx 1.5$, $F = 1$ where climate damages (middle panel) are still relatively low, while other trajectories will approach the less desirable focus point (spiralling steady state) at $C \approx 2.8$, $F = 0.35$ where climate damages are relatively high. Depending on whether countries adjust their time preferences slowly (top panel) or fast (bottom), that focus point is either a stable attractor catching most trajectories that come near it (top) or an unstable repeller which many trajectories have to compass to approach the desirable state after a long transient detour of high damages (bottom). Blue lines show the average development represented by two ordinary differential equations (see Appendix A for details), red lines are the corresponding nullclines (thin: $\dot{F} = 0$, thick: $\dot{C} = 0$), and their other intersection at $C \approx 2$, $F \approx 0.6$ is a saddle point. Parameters: $E_0 = 1.6$, $c = 1$, $r = 0.45$, $l = 0.2$ (top) or 1.3 (bottom), $\gamma = 1.1$, $\mu = 2$, $\sigma = 1$, $\beta = 0.1$, $\alpha = 0.5$, $G = 2$, $N = 50$, $p_0 = 0.5$, $q = 3$.

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

5 Conclusions

In this article, we have presented a taxonomy of processes and co-evolutionary interactions in models of World-Earth systems (i.e. planetary-scale social-ecological systems). For reasons of compactness and compatibility with existing research fields and methodologies we have proposed three taxa for modelled subsystems, and furthermore described a classification of modelled interactions between subsystems into nine taxa. We have illustrated the clarity that this taxonomic framework confers, using a stylised model of social-ecological co-evolutionary dynamics on a planetary scale that includes explicitly socio-cultural processes and feedbacks.

We argue that a relatively simple taxonomy is important for stimulating the discourse on conceptualisations of World-Earth systems. It can help with operational model development as is illustrated by the work reported in the companion paper (Donges et al., 2020). The proposed taxonomy can also help in interdisciplinary communication, model critique, and potentially even participatory modelling processes by providing an organisational scheme and a shared vocabulary to refer to the different components that need to be brought together. However, we acknowledge that alternative, more detailed taxonomies can be beneficial in more specialised settings, e.g. ecological processes are now subsumed in the biophysical taxon, but it may be useful to distinguish them from the geophysical for a clearer understanding of interactions with the socio-metabolic taxon. In other contexts, it may be useful to establish a socio-epistemic taxon separate from the socio-cultural taxon for describing subsystems, processes and interactions involving, for example, symbolic representations and transformations of knowledge through science and technology (Renn, 2018). Along these lines, our framework may be helpful as a blueprint for constructing such alternative, possibly more detailed taxonomies.

Throughout the paper, we have illustrated the taxonomic framework using examples of subsystems, processes and interactions that are already represented in mathematical and computer simulation models in various disciplines. We have not attempted to provide a comprehensive classification of all such modelling components that would be relevant for capturing future trajectories of World-Earth systems in the Anthropocene. Neither have we addressed dynamics beyond the reach of current modelling capabilities, such as long-term evolutionary processes acting within the biophysical taxon or broad patterns and singularities in the dynamics of technology, science, art and history (Turchin, 2008). But we have shown the merits of epistemological pluralism, to enable productive dialogue and interaction between the diversity of World modelling approaches and the biophysical Earth representations that exist and that have agency in a Latourian sense, e.g. through the IPCC processes.

Applying the proposed taxonomy reveals relevant directions in the future development of models of global change to appropriately represent the dynamics of up to planetary-scale social-ecological systems in the Anthropocene. Regarding the sticky problem of representing causality in such a complex system, every possible contributory model is a Pandora's box out of which theoretical controversies and cross-disciplinary battles emerge. The taxonomy outlined here at least partly illuminates what is in this box, making it easier to have more open discussions among modellers about their theories and hypotheses about causality.

While current Earth System Models focus exclusively on representing biophysical subsystems and their interactions and Integrated Assessment Models capitalise on those in the socio-metabolic taxon, socio-cultural subsystems and processes such

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

as the dynamics of opinions and social networks, behaviours, values and institutions and their feedbacks to biophysical and socio-metabolic subsystems remain largely uncovered in planetary-scale models of global change. Integrating these decisive dynamics in World-Earth Models is a challenging, but highly promising research programme (Schellnhuber, 1998, 1999; Steffen et al., 2020) comparable to the development of biophysical Earth system science in the past decades following the foundational blueprints of Bretherton et al. (1986, 1988). We use the copan:DISCOUNT model to demonstrate the value of the taxonomy for tracing how dynamics and feedbacks loop through different taxa, enabling better model design and communication about path-breaking approaches to World-Earth modelling. Following this track will help to develop models that go beyond a climate-driven view of global change and to bridge the “divide" that keeps being spotlighted as the problematic hyphen in prevalent social-ecological/human-nature/etc system concepts. It will also contribute to a deeper understanding of the functioning of complex World-Earth systems machinery in the Anthropocene. By supporting the development and discussion of new family of models, and not by pushing for a rigid and universalising model of everything, applying the taxonomy promises to yield important insights on well-designed policy interventions to foster global sustainability transformation, build World-Earth resilience and avoid social-ecological collapse.

Acknowledgements. This work has been carried out within the framework of PIK’s project on Coevolutionary Pathways in the Earth system (COPAN). It was supported by the Stordalen Foundation via the Planetary Boundary Research Network (PB.net), the Earth League’s EarthDoc programme, the Leibniz Association (project DOMINOES), the Federal Ministry for Education and Research (BMBF, project GLUES), Humboldt University, Berlin, the Swedish Research Council Formas (Project Grant 2014-589), a core grant to the Stockholm Resilience Centre by Mistra, the Heinrich Böll Foundation, the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP/2007-2013)/ERC grant agreement no. 283950 SES-LINK) and under the European Union’s Horizon 2020 Research and Innovation Programme (ERC grant agreement No. 682472 – MUSES and ERC grant agreement No. 743080 ERA). We are thankful to the participants of the LOOPS 2014 workshop on “Closing the loop – Towards co-evolutionary modelling of global society-environment interactions", held at Kloster Chorin, Germany and the LOOPS 2015 workshop on “From limits to growth to planetary boundaries: defining the safe and just space for humanity" held in New Forest, Southampton, UK for inspiring discussions that sparked this paper. Ilona M. Otto, Marc Wiedermann and Finn Müller-Hansen are acknowledged for helpful comments on ideas presented in this paper. We are grateful for the excellent quality of creative stimulus prompted by the comments of Carsten Herrmann-Pillath, Birgit Müller and an anonymous reviewer.

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

References

Anderies, J. M., Carpenter, S., Steffen, W., and Rockström, J.: The topology of non-linear global carbon dynamics: from tipping points to planetary boundaries, Environmental Research Letters, 8, 044 048, 2013.

Arneth, A., Brown, C., and Rounsevell, M.: Global models of human decision-making for land-based mitigation and adaptation assessment, Nature Climate Change, 4, 550–557, 2014.

Arrow, K. J., Cropper, M. L., Gollier, C., Groom, B., Heal, G. M., Newell, R. G., Nordhaus, W. D., Pindyck, R. S., Pizer, W. A., Portney, P. R., Sterner, T., Tol, R. S. J., and Weitzman, M. L.: How Should Benefits and Costs Be Discounted in an Intergenerational Context?, 2013.

Barfuss, W., Donges, J. F., Wiedermann, M., and Lucht, W.: Sustainable use of renewable resources in a stylized social–ecological network model under heterogeneous resource distribution, Earth System Dynamics, 8, 255, 2017.

Barfuss, W., Donges, J. F., Lade, S. J., and Kurths, J.: When optimization for governing human-environment tipping elements is neither sustainable nor safe, Nature communications, 9, 1–10, 2018.

Barfuss, W., Donges, J. F., Vasconcelos, V. V., Kurths, J., and Levin, S. A.: Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse, Proceedings of the National Academy of Sciences, 117, 12 915–12 922, 2020.

Barrett, S.: Self-enforcing international environmental agreements, Oxford Economic Papers, 1994.

Barros, V., Field, C., Dokken, D., Mastrandrea, M., Mach, K., Bilir, T., Chatterjee, M., Ebi, K., Estrada, Y., Genova, R., Girma, B., Kissel, E., Levy, A., MacCracken, S., Mastrandrea, P., and White, L., eds.: Climate Change 2014: Impacts, Adaptation, and Vulnerability. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2014.

Beckage, B., Gross, L., Lacasse, K., Carr, E., Metcalf, S., Winter, J., Howe, P., Fefferman, N., Franck, T., Zia, A., Kinzig, A., and Hoffman, F.: Linking models of human behaviour and climate alters projected climate change, Nature Climate Change, 8, 79–84, 2018.

Bienabe, E. and Hearne, R. R.: Public preferences for biodiversity conservation and scenic beauty within a framework of environmental services payments, Forest Policy and Economics, 9, 335–348, 2006.

Biggs, R., Schlüter, M., Biggs, D., Bohensky, E. L., BurnSilver, S., Cundill, G., Dakos, V., Daw, T. M., Evans, L. S., Kotschy, K., et al.: Toward principles for enhancing the resilience of ecosystem services, Annual Review of Environment and Resources, 37, 421–448, 2012.

Boysen, L. R., Lucht, W., and Gerten, D.: Trade-offs for food production, nature conservation and climate limit the terrestrial carbon dioxide removal potential, Global Change Biology, 23, 4303–4317, 2017.

Bretherton, F. P. et al.: Earth System Science. Overview, Tech. rep., National Aeronautics and Space Administration, Washington, DC, 1986.

Bretherton, F. P. et al.: Earth System Science: A Closer View, Tech. rep., National Aeronautics and Space Administration, Washington, DC, 1988.

Brondizio, E. S., O'brien, K., Bai, X., Biermann, F., Steffen, W., Berkhout, F., Cudennec, C., Lemos, M. C., Wolfe, A., Palma-Oliveira, J., et al.: Re-conceptualizing the Anthropocene: A call for collaboration, Global Environmental Change, 39, 318–327, 2016.

Brugger, J., Feulner, G., and Petri, S.: Baby, it's cold outside: Climate model simulations of the effects of the asteroid impact at the end of the Cretaceous, Geophysical Research Letters, 44, 419–427, 2017.

Brundtland, G. H.: Report of the World Commission on Environment and Development: Our common future, United Nations, 1987.

Budyko, M. I., Ronov, A. B., and Yanshin, A. L.: History of the Earth's atmosphere, Springer, Berlin, 1987.

Calder, M., Craig, C., Culley, D., de Cani, R., Donnelly, C. A., Douglas, R., Edmonds, B., Gascoigne, J., Gilbert, N., Hargrove, C., et al.: Computational modelling for decision-making: where, why, what, who and how, Royal Society Open Science, 5, 172 096, 2018.

Calvin, K. and Bond-Lamberty, B.: Integrated human-earth system modeling—state of the science and future directions, Environmental Research Letters, 13, 063 006, 2018.

Caminade, C., Kovats, S., Rocklov, J., Tompkins, A. M., Morse, A. P., Colón-González, F. J., Stenlund, H., Martens, P., and Lloyd, S. J.: Impact of climate change on global malaria distribution, Proceedings of the National Academy of Sciences, 111, 3286–3291, 2014.

Castellano, C., Fortunato, S., and Loreto, V.: Statistical physics of social dynamics, Reviews of Modern Physics, 81, 591, 2009.

Charney, J., Quirk, W. J., Chow, S.-h., and Kornfield, J.: A comparative study of the effects of albedo change on drought in semi–arid regions, Journal of the Atmospheric Sciences, 34, 1366–1385, 1977.

Colding, J. and Folke, C.: Social taboos: "invisible" systems of local resource management and biological conservation, Ecological Applications, 11, 584–600, 2001.

Crutzen, P. J.: Geology of mankind, Nature, 415, 23–23, 2002.

Cumming, G. S. and Peterson, G. D.: Unifying research on social–ecological resilience and collapse, Trends in Ecology & Evolution, 32, 695–713, 2017.

Dearing, J. A., Wang, R., Zhang, K., Dyke, J. G., Haberl, H., Hossain, M. S., Langdon, P. G., Lenton, T. M., Raworth, K., Brown, S., et al.: Safe and just operating spaces for regional social-ecological systems, Global Environmental Change, 28, 227–238, 2014.

Di Baldassarre, G., Martinez, F., Kalantari, Z., and Viglione, A.: Drought and flood in the Anthropocene: feedback mechanisms in reservoir operation, Earth System Dynamics, 8, 225–233, 2017.

Donges, J. F., Donner, R. V., Marwan, N., Breitenbach, S. F., Rehfeld, K., and Kurths, J.: Non-linear regime shifts in Holocene Asian monsoon variability: potential impacts on cultural change and migratory patterns, Climate of the Past, 11, 709–741, 2015.

Donges, J. F., Lucht, W., Müller-Hansen, F., and Steffen, W.: The technosphere in Earth System analysis: A coevolutionary perspective, The Anthropocene Review, 4, 23–33, 2017a.

Donges, J. F., Winkelmann, R., Lucht, W., Cornell, S. E., Dyke, J. G., Rockström, J., Heitzig, J., and Schellnhuber, H. J.: Closing the loop: Reconnecting human dynamics to Earth System science, The Anthropocene Review, 4, 151–157, 2017b.

Donges, J. F., Heitzig, J., Barfuss, W., Wiedermann, M., Kassel, J. A., Kittel, T., Kolb, J. J., Kolster, T., Müller-Hansen, F., Otto, I. M., Zimmerer, K. B., and Lucht, W.: Earth system modeling with endogenous and dynamic human societies: the copan:CORE open World–Earth modeling framework, Earth System Dynamics, 11, 395–413, 2020.

Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Farahani, E., Kadner, S., Seyboth, K., Adler, A., Baum, I., Brunner, S., Eickemeier, P., Kriemann, B., Savolainen, J., Schlömer, S., von Stechow, C., Zwickel, T., and Minx, J., eds.: Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2014.

Farmer, J. D. and Foley, D.: The economy needs agent-based modelling, Nature, 460, 685–686, 2009.

Fischer-Kowalski, M.: On the history of industrial metabolism, Perspectives on industrial ecology, 2, 35–45, 2003.

Fischer-Kowalski, M. and Erb, K.-H.: Epistemologische und konzeptuelle Grundlagen der sozialen Ökologie, Mitteilungen der Österreichischen Geographischen Gesellschaft, 148, 33–56, 2006.

Fischer-Kowalski, M. and Haberl, H.: Metabolism and colonization. Modes of production and the physical exchange between societies and nature, Innovation: The European Journal of Social Science Research, 6, 415–442, 1993.

Fischer-Kowalski, M. and Hüttler, W.: Society's metabolism, Journal of Industrial Ecology, 2, 107–136, 1998.

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S., Collins, W., Cox, P., Driouech, F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C., and Rummukainen, M.: Evaluation of Climate Models, book section 9, pp. 741–866, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, doi:10.1017/CBO9781107415324.020, www.climatechange2013.org, 2013.

Flato, G. M.: Earth system models: an overview, Wiley Interdisciplinary Reviews: Climate Change, 2, 783–800, 2011.

Folke, C.: Resilience: The emergence of a perspective for social–ecological systems analyses, Global Environmental Change, 16, 253–267, 2006.

Folke, C., Jansson, Å., Rockström, J., Olsson, P., Carpenter, S. R., Chapin III, F. S., Crépin, A.-S., Daily, G., Danell, K., Ebbesson, J., et al.: Reconnecting to the biosphere, Ambio, 40, 719–738, 2011.

Folke, C., Biggs, R., Norström, A. V., Reyers, B., and Rockström, J.: Social-ecological resilience and biosphere-based sustainability science, Ecology and Society, 21, 2016.

Gabriel, M.: Warum es die Welt nicht gibt, Ullstein, 2013.

Gaines, S. D., White, C., Carr, M. H., and Palumbi, S. R.: Designing marine reserve networks for both conservation and fisheries management, Proceedings of the National Academy of Sciences, 107, 18 286–18 293, 2010.

Ganopolski, A., Winkelmann, R., and Schellnhuber, H. J.: Critical insolation–CO2 relation for diagnosing past and future glacial inception, Nature, 529, 200–203, 2016.

Garrett, T. J.: Long-run evolution of the global economy – Part 2: Hindcasts of innovation and growth, Earth System Dynamics, 6, 673–688, doi:10.5194/esd-6-673-2015, http://www.earth-syst-dynam.net/6/673/2015/, 2015.

Geier, F., Barfuss, W., Wiedermann, M., Kurths, J., and Donges, J. F.: The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources, The European Physical Journal Special Topics, 228, 2357–2369, 2019.

Gerten, D., Schönfeld, M., and Schauberger, B.: On deeper human dimensions in Earth system analysis and modelling, Earth System Dynamics Discussions, 2018, 1–22, 2018.

Gregory, J. M., Jones, C., Cadule, P., and Friedlingstein, P.: Quantifying carbon cycle feedbacks, Journal of Climate, 22, 5232–5250, 2009.

Haff, P.: Technology and human purpose: the problem of solids transport on the earth's surface, Earth System Dynamics, 3, 149–156, 2012.

Haff, P.: Humans and technology in the Anthropocene: Six rules, The Anthropocene Review, 1, 126–136, 2014.

Hamilton, C.: Getting the Anthropocene so wrong, The Anthropocene Review, 2, 102–107, 2015.

Harfoot, M. B., Newbold, T., Tittensor, D. P., Emmott, S., Hutton, J., Lyutsarev, V., Smith, M. J., Scharlemann, J. P., and Purves, D. W.: Emergent global patterns of ecosystem structure and function from a mechanistic general ecosystem model, PLoS biology, 12, e1001 841, 2014.

Heck, V., Donges, J. F., and Lucht, W.: Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth System Dynamics, 7, 783–796, 2016.

Heitzig, J., Kittel, T., Donges, J. F., and Molkentin, N.: Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the Earth system, Earth System Dynamics, 7, 21–50, 2016.

Heitzig, J., Barfuss, W., and Donges, J. F.: A thought experiment on sustainable management of the earth system, Sustainability, 10, 1947, 2018.

Herrmann-Pillath, C.: The case for a new discipline: technosphere science, Ecological Economics, 149, 212–225, 2018.

Herrmann-Pillath, C.: The art of co-creation: An intervention in the philosophy of ecological economics, Ecological Economics, 169, 106 526, 2020.

Jarvis, A. J., Jarvis, S. J., and Hewitt, C. N.: Resource acquisition, distribution and end-use efficiencies and the growth of industrial society, Earth System Dynamics, 6, 689–702, doi:10.5194/esd-6-689-2015, http://www.earth-syst-dynam.net/6/689/2015/, 2015.

Jax, K., Barton, D. N., Chan, K. M., De Groot, R., Doyle, U., Eser, U., Görg, C., Gómez-Baggethun, E., Griewald, Y., Haber, W., et al.: Ecosystem services and ethics, Ecological Economics, 93, 260–268, 2013.

Jentoft, S., van Son, T. C., and Bjørkan, M.: Marine protected areas: a governance system analysis, Human Ecology, 35, 611–622, 2007.

Kates, R. W., Clark, W. C., Corell, R., Hall, J. M., Jaeger, C. C., Lowe, I., McCarthy, J. J., Schellnhuber, H. J., Bolin, B., Dickson, N. M., Faucheux, S., Gallopin, G. C., Grübler, A., Huntley, B., Jäger, J., Jodha, N. S., Kasperson, R. E., Mabogunje, A., Matson, P., Mooney, H., III, B. M., and andUno Svedin, T. O.: Sustainability Science, Science, 292, 641–642, 2001.

Kellie-Smith, O. and Cox, P. M.: Emergent dynamics of the climate–economy system in the Anthropocene, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 369, 868–886, 2011.

Keys, P. W. and Wang-Erlandsson, L.: On the social dynamics of moisture recycling, Earth System Dynamics Discussions, 2017, 1–25, 2017.

Kleidon, A.: Thermodynamic foundations of the Earth system, Cambridge University Press, 2016.

Lade, S. J., Niiranen, S., Hentati-Sundberg, J., Blenckner, T., Boonstra, W. J., Orach, K., Quaas, M. F., Österblom, H., and Schlüter, M.: An empirical model of the Baltic Sea reveals the importance of social dynamics for ecological regime shifts, Proceedings of the National Academy of Sciences, 112, 11 120–11 125, 2015.

Lade, S. J., Donges, J. F., Fetzer, I., Anderies, J. M., Beer, C., Cornell, S. E., Gasser, T., Norberg, J., Richardson, K., Rockström, J., and Steffen, W.: Analytically tractable climate-carbon cycle feedbacks under 21st century anthropogenic forcing, Earth System Dynamics Discussions, 2017, 1–23, 2017a.

Lade, S. J., Haider, L. J., Engström, G., and Schlüter, M.: Resilience offers escape from trapped thinking on poverty alleviation, Science Advances, 3, e1603 043, 2017b.

Latour, B.: Facing Gaia: Eight lectures on the new climatic regime, John Wiley & Sons, 2017.

Lenton, T., Schellnhuber, H., and Szathmary, E.: Climbing the co-evolution ladder, Nature, 431, 913, 2004.

Lenton, T. M. and Latour, B.: Gaia 2.0, Science, 361, 1066–1068, 2018.

Lenton, T. M., Pichler, P.-P., and Weisz, H.: Revolutions in energy input and material cycling in Earth history and human history, Earth System Dynamics, 7, 353–370, 2016.

Leontief, W. W.: Quantitative input and output relations in the economic systems of the United States, The Review of Economic Statistics, 18, 105–125, 1936.

Lewis, S. L. and Maslin, M. A.: Defining the anthropocene, Nature, 519, 171–180, 2015.

Lewis-Beck, M. S. and Ratto, M. C.: Economic voting in Latin America: A general model, Electoral Studies, 32, 489–493, 2013.

Lovelock, J. E.: Geophysiology, the science of Gaia, Reviews of Geophysics, 27, 215–222, 1989.

Lovelock, J. E. and Margulis, L.: Atmospheric homeostasis by and for the biosphere: the Gaia hypothesis, Tellus, 26, 2–10, 1974.

Lucht, W. and Pachauri, R.: The mental component of the Earth system, in: Earth system analysis for sustainability, edited by Schellnhuber, H.-J., Crutzen, P., Clark, W., Claussen, M., and Held, H., Dahlem Workshop Reports, pp. 341–365, Cambridge University Press, 2004.

Lutz, W. and Skirbekk, V.: Low fertility in Europe in a global demographic context, in: Demographic Change and Intergenerational Justice, pp. 3–19, Springer, 2008.

Martin, R. and Schlüter, M.: Combining system dynamics and agent-based modeling to analyze social-ecological interactions—an example from modeling restoration of a shallow lake, Frontiers in Environmental Science, 3, 66, 2015.

Masterson, V., Stedman, R., Enqvist, J., Tengö, M., Giusti, M., Wahl, D., and Svedin, U.: The contribution of sense of place to social-ecological systems research: a review and research agenda, Ecology and Society, 22, 2017.

Mengel, M., Nauels, A., Rogelj, J., and Schleussner, C.-F.: Committed sea-level rise under the Paris Agreement and the legacy of delayed mitigation action, Nature Communications, 9, 601, 2018.

Millennium Ecosystem Assessment: Ecosystems and human well-being, Island Press Washington, DC, 2005.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U.: Network motifs: simple building blocks of complex networks, Science, 298, 824–827, 2002.

Mooney, H. A., Duraiappah, A., and Larigauderie, A.: Evolution of natural and social science interactions in global change research programs, Proceedings of the National Academy of Sciences, 110, 3665–3672, 2013.

Morton, T.: Hyperobjects: Philosophy and Ecology after the End of the World, U of Minnesota Press, 2013.

Müller-Hansen, F., Schlüter, M., Mäs, M., Donges, J. F., Kolb, J. J., Thonicke, K., and Heitzig, J.: Towards representing human behavior and decision making in Earth system models–an overview of techniques and approaches, Earth System Dynamics, 8, 977–1007, 2017.

National Research Council: Models in Environmental Regulatory Decision Making, The National Academies Press, Washington, DC, doi:10.17226/11972, https://www.nap.edu/catalog/11972/models-in-environmental-regulatory-decision-making, 2007.

Nelson, G. C., Valin, H., Sands, R. D., Havlík, P., Ahammad, H., Deryng, D., Elliott, J., Fujimori, S., Hasegawa, T., Heyhoe, E., et al.: Climate change effects on agriculture: Economic responses to biophysical shocks, Proceedings of the National Academy of Sciences, 111, 3274–3279, 2014.

Nitzbon, J., Heitzig, J., and Parlitz, U.: Sustainability, collapse and oscillations in a simple World-Earth model, Environmental Research Letters, 12, 074 020, 2017.

Nordhaus, W. D.: An optimal transition path for controlling greenhouse gases, Science, 258, 1315–1319, 1992.

Nordhaus, W. D.: Revisiting the social cost of carbon, Proceedings of the National Academy of Sciences, p. 201609244, 2017.

Ostrom, E.: A general framework for analyzing sustainability of social-ecological systems, Science, 325, 419–422, 2009.

Ostrom, E., Janssen, M. A., and Anderies, J. M.: Going beyond panaceas, Proceedings of the National Academy of Sciences, 104, 15 176–15 178, 2007.

Otto, C., Willner, S. N., Wenz, L., Frieler, K., and Levermann, A.: Modeling loss-propagation in the global supply network: The dynamic agent-based model acclimate, Journal of Economic Dynamics and Control, 83, 232–269, 2017.

Otto, I. M., Donges, J. F., Cremades, R., Bhowmik, A., Hewitt, R. J., Lucht, W., Rockström, J., Allerberger, F., McCaffrey, M., Doe, S. S., et al.: Social tipping dynamics for stabilizing Earth's climate by 2050, Proceedings of the National Academy of Sciences, 117, 2354–2365, 2020a.

Otto, I. M., Wiedermann, M., Cremades, R., Donges, J. F., Auer, C., and Lucht, W.: Human agency in the anthropocene, Ecological Economics, 167, 106 463, 2020b.

Perman, R.: Natural resource and environmental economics, Pearson Education, 2003.

Phalan, B. T.: What have we learned from the land sparing-sharing model?, Sustainability, 10, 1760, 2018.

Purves, D., Scharlemann, J. P., Harfoot, M., Newbold, T., Tittensor, D. P., Hutton, J., and Emmott, S.: Ecosystems: time to model all life on Earth, Nature, 493, 295, 2013.

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

Raworth, K.: A safe and just space for humanity: can we live within the doughnut, Oxfam Policy and Practice: Climate Change and Resilience, 8, 1–26, 2012.

Reichler, T. and Kim, J.: How well do coupled models simulate today's climate?, Bulletin of the American Meteorological Society, 89, 303–311, 2008.

Renn, J.: The Evolution of Knowledge: Rethinking Science in the Anthropocene, HoST-Journal of History of Science and Technology, 12, 1–22, 2018.

Robinson, D. T., Di Vittorio, A., Alexander, P., Arneth, A., Barton, C. M., Brown, D. G., Kettner, A., Lemmen, C., O'Neill, B. C., Janssen, M., Pugh, T. A. M., Rabin, S. S., Rounsevell, M., Syvitski, J. P., Ullah, I., and Verburg, P. H.: Modelling feedbacks between human and natural processes in the land system, Earth System Dynamics Discussions, 2017, 1–47, 2017.

Rocha, J. C., Peterson, G. D., and Biggs, R.: Regime shifts in the Anthropocene: drivers, risks, and resilience, PLoS One, 10, e0134 639, 2015.

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E. F., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., et al.: A safe operating space for humanity, Nature, 461, 472–475, 2009a.

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin III, F. S., Lambin, E., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., et al.: Planetary Boundaries: Exploring the Safe Operating Space for Humanity, Ecology & society, 14, 32, 2009b.

Rockström, J., Gaffney, O., Rogelj, J., Meinshausen, M., Nakicenovic, N., and Schellnhuber, H. J.: A roadmap for rapid decarbonization, Science, 355, 1269–1271, 2017.

Rounsevell, M., Arneth, A., Alexander, P., Brown, D., de Noblet-Ducoudré, N., Ellis, E., Finnigan, J., Galvin, K., Grigg, N., Harman, I., et al.: Towards decision-based global land use models for improved understanding of the Earth system, Earth System Dynamics, 5, 117–137, 2014.

Schellnhuber, H.-J.: Discourse: Earth system analysis – The scope of the challenge, in: Earth system analysis: Integrating science for sustainability, edited by Schellnhuber, H.-J. and Wenzel, V., pp. 3–195, Springer, Berlin, 1998.

Schellnhuber, H. J.: Earth system analysis and the second Copernican revolution, Nature, 402, C19–C23, 1999.

Schill, C., Anderies, J. M., Lindahl, T., Folke, C., Polasky, S., Cárdenas, J. C., Crépin, A.-S., Janssen, M. A., Norberg, J., and Schlüter, M.: A more dynamic understanding of human behaviour for the Anthropocene, Nature Sustainability, 2, 1075–1082, 2019.

Schlueter, M., McAllister, R., Arlinghaus, R., Bunnefeld, N., Eisenack, K., Hoelker, F., MILNER-GULLAND, E., Müller, B., Nicholson, E., Quaas, M., et al.: New horizons for managing the environment: A review of coupled social-ecological systems modeling, Natural Resource Modeling, 25, 219–272, 2012.

Schlüter, M., Baeza, A., Dressler, G., Frank, K., Groeneveld, J., Jager, W., Janssen, M. A., McAllister, R. R., Müller, B., Orach, K., et al.: A framework for mapping and comparing behavioural theories in models of social-ecological systems, Ecological Economics, 131, 21–35, 2017.

Schneider, S. H., Miller, J. R., Crist, E., and Boston, P. J., eds.: Scientists Debate Gaia: The Next Century, MIT Press, Cambridge, Massachusetts, 2004.

Seitzinger, S. P., Gaffney, O., Brasseur, G., Broadgate, W., Ciais, P., Claussen, M., Erisman, J. W., Kiefer, T., Lancelot, C., Monks, P. S., et al.: International Geosphere–Biosphere Programme and Earth system science: three decades of co-evolution, Anthropocene, 12, 3–16, 2015.

Sitch, S., Smith, B., Prentice, I. C., Arneth, A., Bondeau, A., Cramer, W., Kaplan, J., Levis, S., Lucht, W., Sykes, M. T., et al.: Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model, Global Change Biology, 9, 161–185, 2003.

Steffen, W., Crutzen, P. J., and McNeill, J. R.: The Anthropocene: are humans now overwhelming the great forces of nature, AMBIO: A Journal of the Human Environment, 36, 614–621, 2007.

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., de Vries, W., de Wit, C. A., et al.: Planetary boundaries: Guiding human development on a changing planet, Science, p. 1259855, 2015.

Steffen, W., Rockström, J., Richardson, K., Folke, C., Barnosky, A. D., Cornell, S. E., Crucifix, M., Donges, J. F., Fetzer, I., Lade, S. J., Lenton, T. M., Liverman, D., Scheffer, M., Summerhayes, C., Winkelmann, R., and Schellnhuber, H. J.: Trajectories of the Earth system in the Anthropocene, Proceedings of the National Academy of Sciences USA, 115, 8252–8259, 2018.

Steffen, W., Richardson, K., Rockström, J., Schellnhuber, H. J., Dube, O. P., Dutreuil, S., Lenton, T. M., and Lubchenco, J.: The emergence and evolution of Earth System Science, Nature Reviews Earth & Environment, 1, 54–63, 2020.

Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., eds.: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, doi:10.1017/CBO9781107415324, www.climatechange2013.org, 2013.

Strnad, F. M., Barfuss, W., Donges, J. F., and Heitzig, J.: Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos: An Interdisciplinary Journal of Nonlinear Science, 29, 123 122, 2019.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bulletin of the American Meteorological Society, 93, 485–498, 2012.

Traulsen, A., Semmann, D., Sommerfeld, R. D., Krambeck, H.-J., and Milinski, M.: Human strategy updating in evolutionary games., Proceedings of the National Academy of Sciences of the United States of America, 107, 2962–6, 2010.

Turchin, P.: Arise 'cliodynamics', Nature, 454, 34, 2008.

UNFCCC: Paris Agreement, uNTC XXVII 7.d, 2015.

Van Dijk, J. and Breedveld, P. C.: Simulation of system models containing zero-order causal paths—I. Classification of zero-order causal paths, Journal of the Franklin Institute, 328, 959–979, 1991.

van Vuuren, D. P., Bayer, L. B., Chuwah, C., Ganzeveld, L., Hazeleger, W., van den Hurk, B., Van Noije, T., O'Neill, B., and Strengers, B. J.: A comprehensive view on climate change: coupling of earth system and integrated assessment models, Environmental Research Letters, 7, 024 012, 2012.

van Vuuren, D. P., Lucas, P. L., Häyhä, T., Cornell, S. E., and Stafford-Smith, M.: Horses for courses: analytical tools to explore planetary boundaries, Earth System Dynamics, 7, 267–279, doi:10.5194/esd-7-267-2016, 2016.

Vaughan, N. E. and Lenton, T. M.: A review of climate geoengineering proposals, Climatic Change, 109, 745–790, 2011.

Verburg, P. H., Dearing, J. A., Dyke, J. G., van der Leeuw, S., Seitzinger, S., Steffen, W., and Syvitski, J.: Methods and approaches to modelling the Anthropocene, Global Environmental Change, 39, 328–340, 2016.

Vernadsky, V. I.: The biosphere (An abridged version based on the French edition of 1929), Synergetic Press, London, 1929/1986.

Waters, C. N., Zalasiewicz, J., Summerhayes, C., Barnosky, A. D., Poirier, C., Gałuszka, A., Cearreta, A., Edgeworth, M., Ellis, E. C., Ellis, M., et al.: The Anthropocene is functionally and stratigraphically distinct from the Holocene, Science, 351, aad2622, 2016.

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

Watson, A. J.: Implications of an anthropic model of evolution for emergence of complex life and intelligence, Astrobiology, 8, 175–185, 2008.

Weiss, H. and Bradley, R. S.: What drives societal collapse?, Science, 291, 609–610, 2001.

Wiedermann, M., Donges, J. F., Heitzig, J., Lucht, W., and Kurths, J.: Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Physical Review E, 91, 052 801, 2015.

Williamson, O. E.: Transaction cost economics: how it works; where it is headed, De economist, 146, 23–58, 1998.

Winkelmann, R., Levermann, A., Ridgwell, A., and Caldeira, K.: Combustion of available fossil fuel resources sufficient to eliminate the Antarctic Ice Sheet, Science advances, 1, e1500 589, 2015.

Woroniecki, S., Wendo, H., Brink, E., Islar, M., Krause, T., Vargas, A.-M., and Mahmoud, Y.: Nature unsettled: How knowledge and power shape 'nature-based'approaches to societal challenges, Global Environmental Change, 65, 102 132, 2020.

Yearworth, M. and Cornell, S. E.: Contested modelling: a critical examination of expert modelling in sustainability, Systems Research and Behavioral Science, 33, 45–63, 2016.

Zalasiewicz, J., Waters, C. N., Wolfe, A. P., Barnosky, A. D., Cearreta, A., Edgeworth, M., Ellis, E. C., Fairchild, I. J., Gradstein, F. M., Grinevald, J., et al.: Making the case for a formal Anthropocene Epoch: an analysis of ongoing critiques, Newsletters on Stratigraphy, 50, 205–226, 2017.

Zeebe, R. E. and Zachos, J. C.: Long-term legacy of massive carbon input to the Earth system: Anthropocene versus Eocene, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 371, 20120 006, 2013.

Appendix A: The copan:DISCOUNT model

The illustrative model copan:DISCOUNT simulates the co-evolution of $C \geqslant 0$, the excess global atmospheric carbon stock above an equilibrium value that would be attained for zero GHG emissions, and the fraction $F \in [0, 1]$ of the world's countries that care strongly about their future welfare. While C represents the macroscopic state of nature, F represents the macroscopic state of the global human society.

As the derivation of the model below will show, the time evolution of C and F is eventually given by Eqs. 1 and 2. Their governing parameters are business-as-usual emissions $E_0 > 0$, an abatement cost factor $c > 0$, a carbon uptake rate $r > 0$, a learning rate $\ell > 0$, a damage coefficient $\gamma > 0$, a mean tipping point location $\mu > 0$ and spread $\sigma > 0$, two candidate discount rates $0 < \beta < \alpha < 1$, an economic growth factor $G \geqslant 1$, the total number of countries $N > 0$, a curiosity parameter $0 < p_0 < 1$, and a myopic rationality parameter $q > 0$. The equations are derived by combining a standard emissions game model from the literature on international environmental agreements (Barrett, 1994) with a social imitation dynamics that governs the evolution of the countries' time discounting factors as follows.

A1 Countries, welfare

At each point in continuous time, t, a number of $N > 1$ similar countries, i, choose their individual *abatement levels* (carbon equivalents per time), $a_i(t) \geqslant 0$. Global abatement and carbon emissions per time (an interaction of type MET $\rightarrow$ ENV) are

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

then

$$A(t) = \sum_{i=1}^{N} a_i(t), \qquad E(t) = E_0 - A(t), \tag{A1}$$

where $E_0 > 0$ are global "business-as-usual" emissions.

Country i chooses $a_i(t)$ rationally but myopically, only taking into account its own welfare in the present and in "the future" (after a fixed time interval of, say, fifty years). Its present welfare, $W_i^0(t)$, is given by some business as usual welfare, normalised to unity, minus the costs of emissions reductions (MET → CUL), which are a quadratic function of $a_i(t)$ as usual in stylised models of international environmental agreements (Barrett, 1994),

$$W_i^0(t) = 1 - \frac{a_i(t)^2}{2c/N}, \tag{A2}$$

where $c/N > 0$ is a cost parameter that is normalised with N to make the Nash equilibrium outcome (see below) independent of N.

Country i's "future" welfare (belonging to MET), $W_i^1(t)$, is a higher business-as-usual welfare given by a growth parameter $G > 1$, minus the value of additional damages from climate change caused by the present emissions, which are a linear function of $E(t)$:

$$W_i^1(t) = G - s(C(t))E(t), \tag{A3}$$

where $s(C(t)) > 0$ is a *damage factor* that depends on the current carbon stock (see below). Note that while these additional damages $s(C)E(t)$ caused by the present emissions, total damages will still be a nonlinear function of stock C since the factor $s(C)$ changes with C, representing the presence of tipping points (see below).

A2 Discounting, emissions

Since W_i^1 increases in a_i while W_i^0 decreases, choosing an optimal value for a_i involves a trade-off between present and future welfare, which we assume is done in the usual way by using some current *discount factor* $0 < \delta_i(t) < 1$ (an element of taxon CUL) that measures the relative weight of future welfare in country i's optimisation target ("utility") at time t, $U_i(t)$:

$$U_i(t) = (1 - \delta_i(t))W_i^0(t) + \delta_i(t)W_i^1(t). \tag{A4}$$

For simplicity, we assume that only two different discount factors are possible, $0 < \beta < \alpha < 1$, and call a country with $\delta_i(t) = \alpha$ "patient", so that the state of global society at time t can be summarised by the fraction $F(t)$ of patient countries:

$$F(t) = |\{i : \delta_i(t) = \alpha\}|/N. \tag{A5}$$

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under

Given carbon stock $C(t)$ (ENV) and discount factors $\delta_i(t)$, the countries thus face a simultaneous multi-agent multi-objective optimisation problem, each i trying to optimise their utility

$$\begin{aligned}U_i(t) &= \big(1-\delta_i(t)\big)\left(1-\frac{a_i(t)^2}{2c/N}\right)\\ &\quad + \delta_i(t)\Big(G - s\big(C(t)\big)\Big)\left(E_0 - \sum_{j=1}^{N} a_j(t)\right).\end{aligned} \tag{A6}$$

by choosing $a_i(t)$. As in the literature on international environmental agreements, e.g., Barrett (1994), we assume this is solved by making the choices independently and non-cooperatively, i.e., putting $\partial U_i(t)/\partial a_i(t) = 0$ for all i simultaneously, leading to a system of N equations whose solutions $a_i(t)$ form the Nash equilibrium choices (CUL → CUL),

$$a_i(t) = \frac{c}{N}\frac{\delta_i(t)}{1-\delta_i(t)}s(C(t)), \tag{A7}$$

$$\begin{aligned}U_i(t) &= 1 + \delta_i(t)(G - E_0\, s(C(t)) + c\, s(C(t))^2\phi(F(t)) - 1)\\ &\quad - \frac{c}{2N}\frac{\delta_i(t)^2}{1-\delta_i(t)}s(C(t))^2\end{aligned} \tag{A8}$$

and the aggregate abatement (CUL → MET) and emissions

$$A(t) = s(C(t))\,c\phi(F(t)), \qquad E(t) = E_0 - A(t), \tag{A9}$$

where

$$\phi(F(t)) = F(t)\frac{\alpha}{1-\alpha} + (1-F(t))\frac{\beta}{1-\beta}. \tag{A10}$$

A3 Evolution of discount factors

While economic models treat the discount factor of a country as an exogenous parameter, we assume that the value of δ_i is a social trait that may be changed over time due to the observation of other countries' discount factors and their resulting utility (CUL → CUL). As in many models of the spread of social traits (e.g., Traulsen et al. (2010); Wiedermann et al. (2015)), we assume that each country i may adopt another country j's value of δ (social learning by imitation) and that the probability P for doing so depends on the difference between i and j's current utility, $D_{ij}(t) = U_j(t) - U_i(t)$, in a nonlinear, sigmoid-shaped fashion, with $P(D) \to 0$ for $D \to -\infty$ and $P(D) \to 1$ for $D \to \infty$. The utility difference between a country using α and a country using β is

$$\begin{aligned}D(t) &= [\alpha-\beta](G - E_0 s(C(t)) + cs(C(t))^2\phi(F(t)) - 1)\\ &\quad - \left[\frac{\alpha^2}{1-\alpha} - \frac{\beta^2}{1-\beta}\right]\frac{cs(C(t))^2}{2N}.\end{aligned} \tag{A11}$$

This difference is zero iff the discounting summary statistics $\phi(F(t))$ equals

$$\phi_F(C(t)) := \frac{\frac{\alpha^2}{1-\alpha} - \frac{\beta^2}{1-\beta}}{2N[\alpha-\beta]} + \frac{E_0}{cs(C(t))} - \frac{G-1}{cs(C(t))^2} \tag{A12}$$

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

Since $\alpha > \beta$, we have $D(t) > 0$ iff $\phi(F(t)) < \phi_F(C(t))$, meaning that depending on the stock and the fraction of patient countries, either patience or impatience might be more attractive, so that one can expect interesting learning dynamics.

We assume that at each point in time, each country i independently has a probability rate $\ell > 0$ to perform a "learning step". If i does perform a learning step at time t, it compares its current utility $U_i(t)$ with that of a randomly drawn country j and sets its discount factor $\delta_i(t)$ to the value of $\delta_j(t)$ with a probability given by the generalised logistic function,

$$P(D_{ij}(t)) = \frac{1}{1 + \frac{1-p_0}{p_0}\exp(-\frac{q}{p_0(1-p_0)}D_{ij}(t))}, \tag{A13}$$

where $0 < p_0 < 1$ and $q > 0$ are parameters so that $P(0) = p_0$ and $P'(0) = q$.

The "curiosity" parameter p_0 can be interpreted as a measure of a country's curiosity-driven exploration of a different discount factor without expecting a welfare increase. The larger p_0, the more frequent switches will occur, but in both directions between the two candidate discount rates, mainly generating more variance and fluctuations that can be seen as a form of "noise". The "myopic rationality" parameter q can be interpreted as a measure of a country's rationality, because the probability of switching to the other country's discount rate is higher if the other country has higher welfare (and zero if that is not the case) – but it is a myopic rationality, because the agent only takes its present welfare into account. The larger q, the faster discount factors will converge to the one currently generating the largest welfare.

To get a deterministic evolution that can be represented by an ordinary differential equation, we only track the *expected* fraction $F(t)$ of patient countries, which evolves as

$$\dot{F}(t) = \ell F(t)(1 - F(t))[P(D(t)) - P(-D(t))], \tag{A14}$$

while the actual number of patient countries would follow a stochastic dynamics involving binomial distributions that converges to the above in the statistical limit $N \to \infty$. Note that $\dot{F}(t) = 0$ iff $F(t) \in \{0,1\}$ or $\phi(F(t)) = \phi_F(C(t))$.

A4 Carbon stock, damage factor

For ease of presentation, we drop the denotation of time dependence from here on. We assume that the atmospheric carbon stock evolves according to a simplistic dynamics involving only emissions and carbon uptake by other carbon stocks,

$$\dot{C} = E - rC = E_0 - cs(C)\phi(F) - rC \tag{A15}$$

with a constant *carbon uptake rate* $r > 0$ (ENV → ENV). Note that $\dot{C} = 0$ iff $\phi(F)$ equals

$$\phi_C(C) = \frac{E_0 - rC}{cs(C)}. \tag{A16}$$

In order that $C \geqslant 0$ for all times, we require that $\dot{C} \geqslant 0$ whenever $C = 0$, which is ensured by assuming that the parameters fulfil $E_0 \geqslant c\gamma\exp(-\mu^2/2\sigma^2)\phi_1$ where $\phi_1 = \alpha/(1-\alpha)$.

We further assume that $s(C)$, the value (MET → CUL; ENV → CUL) of the additional damages from climate change (ENV → MET; ENV → CUL) due to a marginal increase in emissions at an existing carbon stock C (ENV → ENV), is a positive function of C that has a unique maximum at some critical stock μ at which small changes in stock lead to large changes in damages due to the presence of tipping points. To approximate a damage function that is a sum of a number of sigmoid-shaped functions representing individual tipping points whose locations and amplitudes are roughly normally distributed, we take $s(C)$ to be Gaussian,

$$s(C) = \gamma \exp(-(C-\mu)^2/2\sigma^2), \tag{A17}$$

with parameters $\gamma > 0$, $\mu > 0$, $\sigma > 0$. This completes our derivation of the two ordinary differential equations for C and F.

A5 Steady states, stability

We can distinguish three types of steady states where $\dot{C} = \dot{F} = 0$.

(1) All countries are impatient, $F = 0$ (which implies $\phi(F) = \phi_0 := \beta/(1-\beta)$), and $(E_0 - rC)/cs(C) = \phi_0$. The latter is equivalent to $c\phi_0\gamma \exp(-(C-\mu)^2/2\sigma^2) = E_0 - rC$ which has generically one or three solutions in C with $C > 0$. If there are three, the middle one is always unstable. The others are stable iff $D < 0$.

(2) All countries are patient, $F = 1$ (which implies $\phi(F) = \phi_1$) and $(E_0 - rC)/cs(C) = \phi_1$. The latter is equivalent to $c\phi_1\gamma \exp(-(C-\mu)^2/2\sigma^2) = E_0 - rC$ which again has generically one or three solutions in C with $C > 0$. Again, if there are three, the middle one is always unstable. Again, the others are stable iff $D < 0$. The possibility of two stable states with $F = 1$, one with a small and one with a large C, indicates that even if all countries eventually become patient, this may happen too slowly to prevent a level of climate change (large A) that makes ambitious mitigation even for patient countries too costly in view of the small amount of climate damages that could then still be avoided.

(3) $0 < F < 1$ and $\phi(F) = \phi_F(C) = \phi_C(C)$. This has at most four different solutions in C with $C > 0$, to each of which corresponds at most one solution in F. We know of no simple conditions for assessing their stability but from our numerical experiments we conjecture that (i) at most one of them is stable, namely the one with the largest C, (ii) its stability depends only on the learning rate ℓ, being stable up to a critical value ℓ^*, then unstable; (iii) For $\ell < \ell^*$, it is a stable focus and the leftmost steady state with $F = 0$ is unstable. Hence at most four stable steady states can exist: at most two with $F = 1$, and either at most two with $F = 0$ or at most one with $F = 0$ plus the stable focus with $0 < F < 1$.

Reproduced from: J. F. Donges et al., Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dyn. Discuss., vol. 2018, pp. 1-30, 2018, doi: 10.5194/esd-2018-27. Published under Creative Commons Attribution License 4.0 (CC BY).

Social tipping processes for sustainability: An analytical framework

Authors

Ricarda Winkelmann[1,2]†, Jonathan F. Donges[1,3*†], E. Keith Smith[4,5*†], Manjana Milkoreit[6†], Christina Eder[4], Jobst Heitzig[7], Alexia Katsanidou[4,8], Marc Wiedermann[7], Nico Wunderling[1,2,9], Timothy M. Lenton[10]*

Affiliations

(1) FutureLab on Earth Resilience in the Anthropocene, Earth System Analysis, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Telegrafenberg A31, 14473 Potsdam, Germany
(2) Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany
(3) Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden
(4) GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany
(5) International Political Economy and Environmental Politics, ETH Zurich, Switzerland
(6) Department of Political Science, Purdue University, 100N University Street, West Lafayette, IN 47906, United States of America
(7) FutureLab on Game Theory and Networks of Interacting Agents, Complexity Science, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Telegrafenberg A31, 14473 Potsdam, Germany
(8) Institute of Sociology and Social Psychology, University of Cologne, Cologne, Germany
(9) Department of Physics, Humboldt University of Berlin, Berlin, Germany
(10) Global Systems Institute, University of Exeter, Exeter, EX4 4QE, United Kingdom
** Corresponding authors: ricarda.winkelmann@pik-potsdam.de, jonathan.donges@pik-potsdam.de, keith.smith@gess.ethz.ch*
† shared lead authorship

Abstract

Societal transformations are necessary to address critical global challenges, such as mitigation of anthropogenic climate change and reaching UN sustainable development goals. Recently, social tipping processes have received increased attention, as they present a form of social change whereby a small change can shift a sensitive social system into a qualitatively different state due to strongly self-amplifying (mathematically positive) feedback mechanisms. Social tipping processes have been suggested as key drivers of sustainability transitions emerging in the fields of technological and energy systems, political mobilization, financial markets and sociocultural norms and behaviors.

Drawing from expert elicitation and comprehensive literature review, we develop a framework to identify and characterize social tipping processes critical to facilitating rapid social transformations. We find that social tipping processes are distinguishable from those of already more widely studied climate and ecological tipping dynamics. In particular, we identify human agency, social-institutional network structures, different spatial and temporal scales and increased complexity as key distinctive features underlying social tipping processes. Building on these characteristics, we propose a formal definition for social tipping processes and filtering criteria for those processes that could be decisive for future trajectories to global sustainability in the Anthropocene. We illustrate this definition with the European political system as an example of potential social tipping processes, highlighting the potential role of the FridaysForFuture movement.

Accordingly, this analytical framework for social tipping processes can be utilized to illuminate mechanisms for necessary transformative climate change mitigation policies and actions.

Keywords

Social tipping dynamics, social change, sustainability, critical states, network structures, FridaysForFuture

MAIN TEXT

1. Introduction

There is growing concern that global climate change is reaching a point where parts of the Earth System are starting to pass damaging climate tipping points (*1*): In particular, part of the West Antarctic Ice Sheet (WAIS) appears to already be collapsing because of irreversible retreat of grounding lines (*2*, *3*) which in turn is expected to trigger loss of the rest of the WAIS (*4*). Other tipping points may be close: A recent systematic scan of Earth system model projections has detected a cluster of abrupt shifts between 1.5 and 2.0°C of global warming (*5*), including a collapse of Labrador Sea convection with far-reaching impacts on human societies. The abrupt degradation of tropical coral reefs is projected to be almost complete if warming reaches 2.0°C (*6*, *7*). The possibility of the global climate tipping to a 'hothouse Earth' state has even been posited (*8*).

Against this backdrop, there is a growing consensus that avoiding crossing undesired climate tipping points requires rapid transformational social change, which may be propelled (intentionally or unintentionally) by triggering social tipping processes (*9*, *10*) or "sensitive intervention points" (*11*, *12*). Examples for such proposed social tipping dynamics include divestment from fossil fuels in financial markets, political mobilization and social norm change, socio-technical innovation (*9–11*, *13*, *14*). Equally, if human societies do not act collectively and decisively, climate change could conceivably trigger undesirable social tipping processes, such as international migration bursts, food system collapse or political revolutions (*15*). Social tipping processes have received recent attention, as they encompass this sort of rapid, transformational system change (*9*, *10*, *13*, *15*).

Here we develop an analytical framework for social tipping processes. Drawing upon expert elicitation and a comprehensive literature review, we find that the mechanisms underlying social tipping processes are categorically different from other forms of tipping, as they uniquely have the capacity for agency, they operate on networked social structures, have different spatial and temporal scales, and a higher degree of complexity. Following these distinctions, we present a definitional framework for identifying social tipping processes for sustainability, where under critical conditions, a small perturbation can induce non-linear systemic change, driven by positive feedback mechanisms and cascading network effects. We adopt this framework to understand potential social tipping dynamics in the European political system, where the *FridaysForFuture* movement (*16*) pushes the system towards criticality, generating the conditions for shifting climate policy regimes into a qualitatively different state.

The proposed framework aims to establish a common terminology to avoid misconceptions, including the notions of agency, criticality as well as the manifestation and intervention time horizons in the context of social tipping. In this way, the framework can serve to connect literatures and science communities working on social tipping, social change, complex contagion dynamics and evidence from behavioral experiments (e.g. *14*, *17*).

2. Background

2.1. Tipping points as social-ecological systems features

We start by reviewing the characterization of tipping points across the natural and social sciences. Over the last 150 years, a suite of concepts and theories describing small changes with large systemic effects has been developed at the intersection of natural and social sciences. More recently, the concepts of tipping points and tipping elements have been broadly adopted by both natural and social scientists working within the field of climate change.

While the concept of 'tipping' originated in the natural sciences (*18*, *19*), social scientists made extensive use of the idea in the 20th century, often without using the terminology of tipping.

Famously, Schelling (*20*), following Grodzins (*21*), developed a theory of tipping processes to explain racial segregation in US neighbourhoods. Granovetter (*22*) modeled collective behavior as a tipping process that depends on passing individual thresholds for participation in riots or strikes. Kuran (*23*) described political revolution in terms of tipping dynamics, while Gould and Eldridge (*24*) distinguish phases of policy change and stability in terms of 'punctuated equilibrium'. Gladwell (*25*) popularised the concept of 'tipping points', exploring contagion effects ("fads and fashions"), sometimes triggered by specific events.

Several recent studies have examined tipping processes within contemporary social systems. Homer-Dixon (*26*) and Battison (*27*) explored the 2008 financial crisis as a tipping phenomenon. Nyborg (*14*, *28*) discussed shifts in norms and attitudes, for example regarding smoking behaviors. Centola (*17*) associated tipping points with the "critical mass phenomenon", wherein 20–30% of a population becoming engaged in an activity can be sufficient to tip the whole society. Similarly, Rockström et al. (*29*) highlighted this so-called Pareto effect in the context of decarbonization transitions. Kopp et al. (*15*) distinguished different social tipping elements within the realm of policy, new technologies, migration and civil conflict that are sensitive to "climate-economic shocks". Here, a tipping element is a system or subsystem that may undergo a tipping process.

Since the mid 1990s, ecologists and social-ecological systems (SES) researchers have also developed an extensive body of research on tipping processes using the terminology of 'regime shifts' and 'critical transitions' (e.g. *30*–*32*). Recognizing the impacts of human development on various ecosystems, this body of work often models ecological regime shifts as a consequence of social drivers. Less attention, however, has been paid to sudden changes in social systems triggered by ecosystem changes.

There is a rich literature on the collapse of past civilizations (e.g. *33*, *34*) and the potential role of tipping points in that (*35*). Recently, Cumming and Peterson (*36*) brought this together with work on ecological regime shifts, proposing a "unifying social-ecological framework" for understanding resilience and collapse. Further, Rocha et al. (*37*) noted that tipping dynamics can be produced by the interactions between climatic, ecological and social regime shifts.

The concept of climate tipping elements introduced by Lenton et al. (*1*) and Schellnhuber (*38*), has been increasingly adopted within Earth and climate sciences. Climate tipping elements are defined as at least sub-continental-scale components of the climate system that can undergo a qualitative change once a critical threshold in a control variable, e.g., global mean temperature, is crossed. Positive feedback mechanisms at the critical threshold drive the system's transition from a previously stable to a qualitatively different state (*1*). Other scholars, e.g., Levermann et al. (*39*), suggest a somewhat narrower definition of climate tipping elements by introducing additional characteristics, such as (limited) reversibility or abruptness. The tipping elements identified so far include biosphere components such as the Amazon rainforest (*40*–*42*) and coral reefs (*6*, *7*), cryosphere components such as the ice-sheets on Greenland and Antarctica (*43*), and large-scale atmospheric or oceanic circulation systems including the Atlantic meridional overturning circulation (*44*, *45*). Their tipping would have far-reaching impacts on the global climate, ecosystems and human societies (e.g. *8*, *46*).

2.2. Social Tipping

In response to the concept of climate tipping points, social scientists are re-engaging with this concept yet again, creating an additional layer of tipping scholarship with an emphasis on the need for and possibility of deliberate tipping of social systems onto novel development pathways towards sustainability (e.g. *11*, *47*). Scholars argue in particular that the rapid, non-linear change of social tipping dynamics might be necessary to speed up societies' responses to climate change, and to achieve the goals of the Paris Agreement. It is this element of acceleration, propelled by positive feedbacks, that makes the concept of tipping particularly interesting. For example, Otto and Donges et al. (*9*) reported expert elicitations identifying social tipping elements relevant for driving rapid decarbonization by 2050. Rapid-paced changes are a distinctive feature potentially differentiating tipping dynamics from many other forms of social change, including incremental (policy or institutional) changes, or more radical (socio-technical) transitions or societal transformations.

Over the last decade, the literature on deliberate transitions and transformations towards sustainability has expanded significantly, exploring the dynamics that lead to the reorganization of social, economic or political systems (e.g. *48*, *49*). In many ways, this literature and the emerging work on social tipping are interested in very similar phenomena: fundamental shifts in the organization of social or social-ecological systems - a movement from one stable state to another - including a change in power relations, resource flows, as well as actor identities, norms and other meanings (*48*). Transformations can be fast, but speed is generally not one of their defining characteristics.

This temporal feature of social tipping points - rapidity of change compared to the system's normal background rate of change - combined with the fact that tipping processes can be triggered by a relatively small disturbance of the system is motivating scholarship on leverage or 'sensitive intervention points', e.g. Farmer et al. (*12*), who identified such potentially high-impact intervention opportunities, e.g., financial disclosure, choosing investments in technology and political mobilization that may be key for triggering decarbonization transitions.

Based on a bibliometric and qualitative review of these various bodies of literature across the natural and social sciences, Milkoreit et al. (*10*) proposed the following general definition of (social) tipping: "the point or threshold at which small quantitative changes in the system trigger a non-linear change process that is driven by system-internal feedback mechanisms and inevitably leads to a qualitatively different state of the system, which is often irreversible." Milkoreit et al. (*10*) further noted there is a need to recognize and identify potential differences between climatic (or ecological) and social tipping processes to gain a deeper understanding of these phenomena.

3. Methods and analytical structure

Given this diverse and nascent field, there is a clear need for consensus as to what defines social tipping processes, as well as an understanding of how these processes are similar and diverge from dynamics in other non-social systems. Further, there are currently limited examples of social tipping elements in the context of sustainability transitions presented within the broader literature (*9*, *12*, *13*, *15*).

Here we explore the characterization of tipping processes within the natural and social sciences, examining how social and climate tipping processes are differently conceptualized. We draw upon a mixed qualitative methodological approach to illuminate these differences and key distinctions. Initially, core differences were identified and discussed via expert elicitation (*50*). A selected group of 25 experts from across the climate and social sciences were invited to take part in an expert elicitation workshop, that focused on identifying a common definition for social tipping processes, as well as the characterization of their dynamics. This workshop was convened in June 2018 in Cologne, Germany. The workshop participants were split into cross-disciplinary breakout groups, to independently identify the dynamics of social tipping processes. Then, each of these groups reported their findings to the broader plenary, for discussion, consolidation, reconciliation and clarification. The process was then repeated for further clarification within the breakout groups. Through this iterative inductive and deductive process, several unique themes and characteristics were identified from the broader set of codes, resulting in the key differences in and definition of social tipping processes presented below.

Drawing upon the differences identified in the expert elicitation workshop, we then review and synthesize the emerging field of social tipping processes, particularly in comparison to the related climate and ecological tipping dynamics. We then draw upon these unique characteristics to develop a common definition for social tipping processes, which we explore using the example of the *FridaysForFuture* student movement.

4. Results

4.1. Key differences between social and climate tipping processes

Social and climate systems' tipping processes exhibit several broad, fundamental differences in their structure and underlying mechanisms: (i) agency is a main causal driver of social tipping processes, (ii) the quality of social networks and associated information exchange provides for specific social change mechanisms not available in non-human systems, (iii) climate and social tipping processes occur at different spatial and temporal scales, and (iv) social tipping dynamics exhibit significantly more complexity than climatic ones.

Agency: The most important characteristic differentiating social from climate tipping processes is the *presence of agency*. While a significant body of work (e.g. *51*), including Latour's actor-network theory (*52*), addresses different forms and effects of non-human or more-than-human agency, here, we focus on a more narrow understanding of agency that is based on consciousness and cognitive processes such as foresight, planning, normative-principled and strategic thinking, that allow human beings to purposefully affect their environment on multiple temporal and spatial scales. While humans have a generally poor track record of utilizing their agentic capacities especially with regard to shaping the future (e.g. *53–55*), they appear unique in their capacity to transcend current realities with their decisions.

Agency in this more narrow sense can be understood as the human capacity to exercise free will, to make decisions and consciously chart a path of action (individually or collectively) that shapes future life events and the environment (*56*). The notion of intentionality inherent in the idea of agency implies that human actors are not only able to adapt to changes in their environment, but also deliberately create such changes. Non-human life forms can also be engaged in deliberate changes of their environment (e.g., beavers building dams), but the cognitive quality of these actions differs from those of humans, which can be based on different forms of knowledge and meaning about the world, moral norms and principles, or ideas about desirable futures. Agency allows individuals and societies to be proactive rather than merely responsive in their relationships with other humans or the environment through planning, goal setting and strategic decision-making, which links decisions and behaviors in the present with consequences and realities in the (distant) future (*57*).

Governance scholars address this social-cognitive capacity for forethought and goal-pursuit in terms of anticipation (*58*) and imagination (*10*), which can be tied to a set of futuring methods (*59*, *60*). The ability to anticipate and imagine futures enables humans and their societies (*53*, *54*) – as opposed to animal communities or ecosystems – to transcend the present and shape the future according to our values and goals (*61*), possibly increasing the prospects for human survival in times of fast and significant environmental change (*56*, *62*). Although this ability has been underutilized in the past, especially in the context of responding to climate change (*63*), it is a crucial dimension of the human repertoire of tools to create change and to ensure its long-term well-being.

Agency interacts with many of the additional differentiating characteristics we identify below in important ways. For example, agency plays a role in the creation of social networks, institutions and meaning, i.e., the production of the structures of social systems. These network structures in turn enable and constrain agency (e.g. *64*, *65*).

Physical climate tipping elements, such as ice sheets or ocean circulations, lack that ability to intentionally act and adapt. However, the adaptive capacity of ecosystems can be interpreted as a form of non-human agency and learning mechanism (*66*), see also Supplementary Information S2. While scholarship on non-human agency, including that of animals, inanimate objects, landscape features or ecosystems (e.g. *67*, *68*) might expand our understanding of agency, the cognitive abilities that characterize human agency, especially long-term and strategic thinking, do not exist in the non-human or inanimate worlds.

Social networks: Understanding the *nature of social networks* is crucial for studying social tipping. While both natural (including physical and ecological) and social systems can be structurally characterized as networks and studied using a network science approach (*69*), social systems differ from natural systems in the quality of the networks' nodes and interconnections and the processes and dynamics facilitated and impacted by these particular network characteristics. Social systems feature additional network levels of information transmission (cultural and symbolic) that are largely restricted to human societies compared to natural systems (*70*).

Network qualities unique to social systems:
Networks in social and natural systems share various commonalities such as the existence of fundamental nodes and links (*69*). In contrast to most natural systems, however, social networks have the capacity to intentionally generate new nodes, which include socially constructed entities such as organizations and movements (*71*). New nodes can be created through cultural, political or legal means, as can the rules for their interactions with other existing nodes. Social system nodes are unique in that they have richer cognitive realities, particularly agency and forethought. These nodes often have conflicting vested interests, which may be more short-sighted than future oriented.

Relationships in social networks can consist of shared meanings – especially norms, identities and other ideas – and a vast variety of cultural, economic and political relationships (e.g., employment, citizenship), all of which are not as pronounced or non-existent in less complex human societies and nature. Hence, social network links are more diverse than links in natural systems and enable different kinds of network processes. For example, links between nodes in social networks are not necessarily dependent on physical co-presence, due to technologically enabled connections or the presence of more abstract interrelations such as shared norms, values or interpersonal relationships.

Network processes:
Social network dynamics can be of a purely ideational nature (e.g., the subject of the study of opinion and belief dynamics), but also involve material changes (e.g., resource extraction, movement and transformation for economic purpose). Markets are unique social networks, involving both ideational and material network processes. In the Anthropocene, the intensity and speed of socially networked interaction has increased dramatically, largely due to new media, digitalization, more efficient means of transportation, lower travel costs, and overall increased mobility, which is likely to increase spreading rates, while at the same time affecting the stability of the network itself (*72–74*).

Generally, social tipping can either occur on a given network (e.g., through spreading dynamics changing the state of nodes (*75*) or change the network structure itself (see Figure 1). The structural network changes generated by social tipping processes include transitions from centralistic or hierarchical to more polycentric (neuromorphic) structures in urban systems, energy distribution and generation networks (*76*, *77*). Structural changes can manifest on large and small-scale spatial networks across multiple social structure levels. In order to capture these network tipping processes, quantifiers from complex network theory such as modularity, degree distribution, centrality or clustering can be used (*69*).

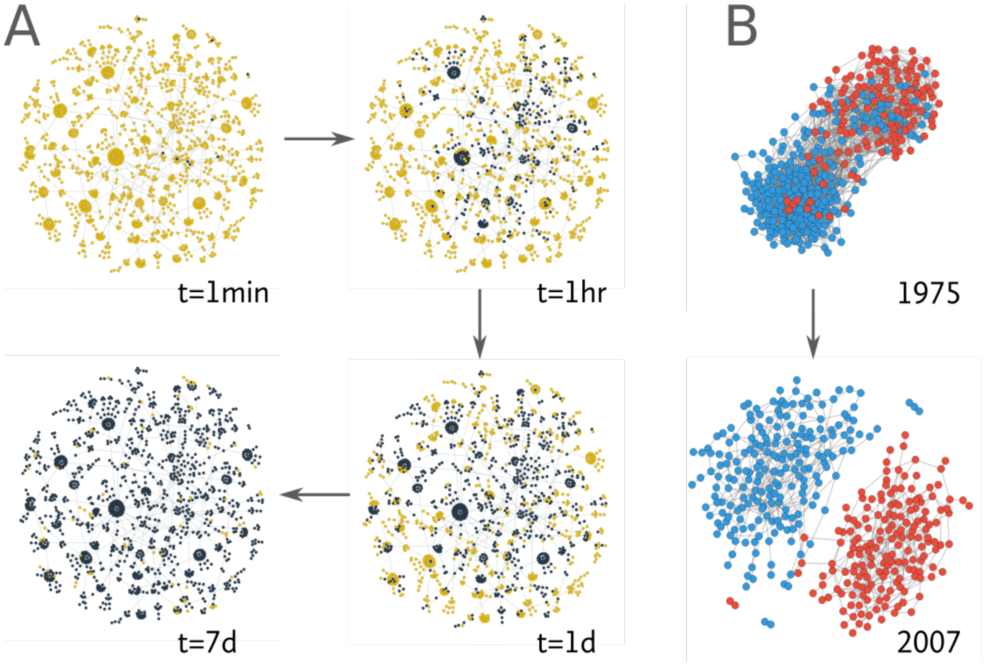


Figure 1: Two types of social tipping in a complex network. *(A) Social tipping can on the one hand be characterized by a contagion process where initially only a few nodes exhibit a certain property that then spreads through a large portion of the network. (B) On the other hand social tipping may also qualitatively alter the entire network structure from, e.g., a state with closely entangled nodes of different states to an almost or full disintegration of the network in smaller disjoint groups. The example in (A) shows the spread of an avatar among users in an online virtual world over the course of one week after it was first introduced by a small number of users* ***(78)****. Nodes represent users and links represent the imitation of the avatar from one user to another. Yellow nodes denote users that have not picked up the avatar, while black nodes indicate those that did. (B) The upper network shows the members of the House of Representatives in the 94th United States Congress (January 3, 1975 to January 3, 1977). Node colors indicate different party membership and links between nodes are drawn if the corresponding members agree on 66% of all votes in the considered two-year period. The lower network shows the same for the 110th United States Congress (January 3, 2007, to January 3, 2009). The transition from a closely entangled to an almost fragmented topology indicates a polarisation between Democratic and Republican Party members over time (16).*

Temporal and spatial scales: Scales can differ greatly between social tipping and climate tipping processes and are more ephemeral for social tipping than for climate tipping.

Temporally, tipping in social systems manifests more commonly on the scale of *months to decades*, while for the climate tipping elements range from *years to millennia*. Human actors tend to focus on more short-term consequences or outcomes, as complex issues (such as climate change) with longer timeframes are often harder to assess (*79*). Within social systems, fund manager performance is evaluated quarterly, politicians often think in electoral cycles, business operates with annual or five-year forecasts, while individual practices and dispositions are constantly evaluated and reevaluated (*80–82*). In natural systems, however, it might take decades, centuries or even millennia for outcomes of change processes to become detectable (see Figure 2).

Both social and climate tipping elements can be ordered spatially (*1*, *39*, *83*), although social tipping elements cannot always be precisely located geographically. Social scientists and economists have long grouped systems and processes as existing on the macro-, meso- and micro-levels (or some variation thereof), whereby some social systems (e.g., financial markets, political systems, technologies) consist of interdependent subsystems existing on multiple spatial levels.

Social tipping processes can also display spatial-temporal *ephemerality*. While climate tipping elements have a known spatial extent and dimensionality (with often a comparable extent in latitude and longitude and a generally much smaller extent in altitude) and have persisted in their current stable state for thousands (if not millions) of years, social tipping processes do not have a spatial extent or effective dimensionality that is known ex-ante and they can emerge (move into a critical state) and disappear (move out of a critical state) over time.

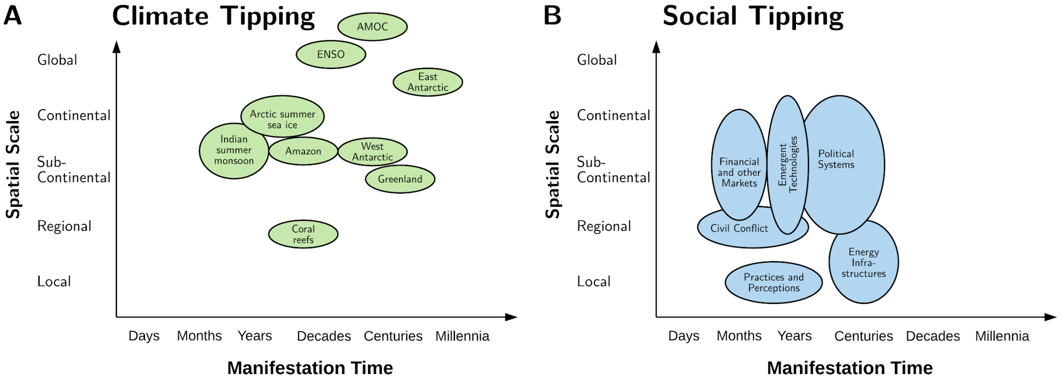


Figure 2: Examples of spatial and temporal scales for climate and social tipping elements. *Example climate tipping elements are broadly compiled from Lenton et al. (1), Levermann et al. (39), and Schellnhuber et al. (83). Social tipping elements are broadly compiled from Kopp et al. (15), Farmer et al. (12), Otto and Donges et al. (9), Hsiang (84), Tabara (11) and Lenton (13).*

Complexity: Social tipping processes occur in complex *adaptive* systems (*85–87*) as opposed to the complex but non-adaptive physical climate system. As such they can exhibit comparatively *greater complexity* in the (i) drivers, (ii) mechanisms and (iii) resulting pathways of social tipping processes, as well as the aforementioned ephemerality in their spatial-temporal manifestations, including a potentially fractal and varying dimensionality and a more complex interaction topology (*88*, *89*).

Social tipping processes can rarely be linked to a single common control parameter, such as is the case with global mean temperature in climate tipping dynamics. For most of the climate tipping elements like the ice sheets or the Atlantic meridional overturning circulation, the control variables such as local air temperature, precipitation or ocean heat transport, can often be translated or downscaled into changes in global mean temperature as one common driver (*1*, *38*). However, for social tipping processes, multiple, interrelated factors are often identified as forcing the critical transition. For example, shifts in social norms regarding smoking (*14*) can be linked to several, entwined factors, such as policies, taxation, advertising and communication, social feedbacks (e.g., via normative conformity), or individual preference changes. Centola et al. (*17*) show that tipping in social convention is possibly explained by a single parameter: the size of the committed minority). At larger scales, the collapse of complex civilizations has been linked to multiple interacting causes, and whilst disagreement abounds over the balance of causes in particular cases, there is general agreement that multiple factors were at play (*33*). This kind of causality – multiple interacting, distributed causes across varying scales – are a key characteristic of complex systems (*90*), contrasting starkly with conventional notions of causality involving bivariate relationships (one cause and one effect).

Further, due to their potential for agency and adaptive plasticity, social systems are open to a larger number of mechanisms that could cause a tipping process and various pathways of change that a tipping process could follow towards a greater number of potentially stable post-tipping states (*91*). Climate tipping processes are often modeled as bi- or multistable, where the directional outcomes of forcing are to some extent known or knowable, e.g., based on paleoclimatic data and process-based Earth system modelling. Given a specific forcing change, one can predict in what state the element will restabilize as well as the "net" effects of the tipping process on larger Earth systems. Based on this understanding, the tipping of climate system elements is generally perceived as undesirable and often as part of pushing the Earth system out of the "safe operating space for humanity" (*92*, *93*).

In contrast, for social systems, it is often unclear what a final stable state of the system will look like, or even whether the changes resulting from a tipping process will be normatively considered "positive" or "negative". As Clark and Harley (*94*) point out, the characteristics of complex-adaptive social systems, including the diversity of actors and elements and the different outcomes generated by local and global interactions, imply that the development pathways of these systems are less predictable. Further, a social tipping process can generate new and destroy existing actor types (e.g., identities, institutions) and their behaviors. Cross-scale dynamics and local differences are important to understand the emergent system structure and change dynamics, but predictive capacities, e.g., regarding the timing of a social tipping point or the boundaries between different stable states, do not yet exist (*94*). Hence, the term 'managing transitions' is less useful than the idea of navigating a transformation pathway.

The political nature of social change processes (*95*) – different actors within a social community pursuing different, sometimes opposing, interests and visions for a reorganization of a social system while bringing to bear different resources and strategies – further exacerbate this situation. Actors can deliberately generate new feedback dynamics that support or slow change, even after a tipping point has been passed, and they can actively work to adjust the direction of change.

4.2. Proposed definition of social tipping processes

From the discussion above, it follows that a definition of social tipping process should take a micro-perspective and incorporate network effects and agency in addition to common tipping characteristics already explored in the review by Milkoreit et al. (*10*). It should also describe the timing aspects sufficiently well to understand possibilities for intervention, similar to what Lenton et al. (*1*) suggested for climate tipping elements. Hence we propose the following definition of the various terms relevant for studying social tipping processes (see Supplementary Material S1 for a more formal mathematical definition suggested for use in simulation modelling and data analysis that is consistent with what we put forward here):

> ***Definitions:*** *A '<u>social system</u>' can be described as a network consisting of social agents (or subsystems) embedded within a social-ecological '<u>environment</u>'. Such a social system is called a '<u>social tipping element</u>' if under certain ('<u>critical</u>') conditions, small changes in the system or its environment can lead to a qualitative (macroscopic) change, typically via cascading network effects such as complex contagion and positive feedback mechanisms. Agency is involved in moving the system towards criticality, creating small disturbances and generating network effects. By this definition, near the critical condition the stability of the*

social tipping element is low. The resulting change process is called the 'tipping process. The time it takes for this change to manifest is the 'manifestation time'.[1]

If a tipping element *is* already in a critical condition, where the stability of its current state is low, there may be a time window during which an agential intervention might *prevent* an unwanted tipping process by moving the system into an uncritical condition (see also SI text S1). Alternatively, if a tipping element *is not* already in a critical condition, there may be a time window during which some intervention might move it into a critical condition in order to *bring about* a desired tipping process.

The small change triggering the tipping process could be either (i) a localized modification of the network structure (e.g., a change on the level of single nodes, small groups of nodes or links) or of the state of agents or subsystems, (ii) small changes of macroscopic parameters or properties, or (iii) small external perturbations or shocks. We deliberately do *not* require the trigger to be a *single* driving parameter. This is because we expect that a social tipping process could be triggered by a *combination* of causes rather than a single cause. Furthermore, a social tipping element may be tipped by several *different* combinations of causes. Consequently, for social tipping elements we cannot always expect at this point to identify a common aggregate indicator (such as global mean temperature in the case of climatic tipping elements) and a well-defined 'threshold' for this indicator at which the system will tip (see also the discussion on complexity above).

Note that social tipping as defined here is a unique form of social change, e.g., distinct from climate economic shocks (*15*) and more specific than socio-technical transitions (*96*, *97*). Further, social tipping also denotes a shift to a qualitatively different state, and such, is different from standard business cycles or causes of seasonality. As such, social tipping presents a particular process of social change, where a system undergoes a transformation from one qualitatively different state to another, after being in a more critical state and affected by a potentially small triggering event.

4.3. Filtering criteria

We propose several filtering criteria to focus on social tipping processes (i) that have the potential to be relevant to global sustainability in future Earth system tractories and (ii) where human interventions can occur within a pertinent *intervention time horizon* on the order of decades and will have consequences within a *political/ethical time horizon* on the order of hundreds of years.

(i) Relevance of social tipping for global sustainability

The social tipping process can impact a wide array of social systems, such as technological or energy systems, political mobilization, financial markets and sociocultural norms. We consider social tipping processes to be relevant here that have an impact on the biophysical Earth system or on macro-scale social systems. The qualitative change in a 'relevant' social tipping process significantly affects the future state of the Earth system in the Anthropocene directly or indirectly through interactions with other social tipping processes. Relevance can hence be defined in terms of impacts on biophysical Earth system properties such as global mean temperature, biosphere integrity or other planetary boundary dimensions. For example, tipping dynamics to a political system could result in policy regime changes, affecting substantial reductions in greenhouse gas emissions (*9*, *12*). Furthermore, we consider social tipping processes that have relevant impacts on macro-social systems and can be triggered by changes in the same biophysical Earth systems, for example, mass migration due to climate impacts (*84*, *98*).

[1] This is analogous to the 'transition time' in Lenton et al. (*1*) . We avoid the term 'tipping *point*' in this definition since some of the literature uses it to refer to a point in time while some of the literature uses it to refer to a certain state of the system or its environment.

(ii) Intervention and ethical time horizons

We are interested in potential social tipping processes in which humans have the agency to substantively intervene. For example, such interventions could be via technological or physical capacities of agential or structural actors. This therefore places emphasis on human intervention, such as decreasing the likelihood of extreme weather events via mitigation efforts, or triggering socio-technological changes towards decarbonization. We define intervention and ethical time horizons as follows:

Intervention time horizon
Human agency interferes with a social tipping element, such that decisions and actions taken between now and an 'intervention time horizon' could influence whether (or not) the system tips. We suggest to consider only social tipping processes with an intervention time on the order of 10 years (*9*), which arguably presents a practical limit of human forethought (*99*) and of future-oriented political agency. For example, international governance efforts for global sustainability challenges, such as the ozone regime or the Sustainable Development Goals, tend to work with similar time horizons. Similarly, social tipping processes for rapid decarbonization to meet the Paris climate agreement would have to be triggered within the next few years (*9*), with ambitious emissions reduction roadmaps aiming for peak greenhouse gas emissions in 2020 (*29*, *100*). The intervention time horizon is analogous to the 'political time horizon' defined for climate tipping elements in Lenton et al. (*1*).

Ethical time horizon
The time to observe these relevant consequences should lie within an 'ethical time horizon'. This recognizes that consequences manifesting too far in the future are not relevant to the current discourse on how contemporary societies impact Earth systems. Such an ethical time horizon could consider only social tipping processes which can have relevant *consequences within the next centuries* at most, corresponding to an upper life expectancy of the next generations of children born.

4.4. Example of a potential social tipping process: European Climate Change Policy Dynamics Europe and FridaysForFuture

Currently, international climate policies, including those of the European Union (EU) are insufficient to meet the +1.5°C or +2°C goals of the Paris Agreement (*101*). While European policy makers presume to lead global mitigation efforts and characterize their actions as ambitious (*102*, *103*), actual policy measures and proposals have been lagging behind this aspiration (*104*). EU countries emit about a tenth of the world's emissions, and a policy change towards more rapid decarbonization would not only have significant direct impacts on the climate system, but likely have indirect effects on the policies of other major emitters. But what kinds of sociopolitical processes can lead to these necessary changes? Could such changes result from social tipping dynamics?

Public opinion is a crucial factor in policy formation, where the public can be understood as a "thermostat" signalling what is politically feasible (*105*, *106*). Shifts in public opinion can punctuate previously stable and 'sticky' institutions, leading to policy change (*107*). Increased activism and public concern regarding climate change can generate new coalitions, or shift the priorities of existing ones (*108*, *109*). Here we examine the European political system as an example of and how social tipping processes could be triggered as a result of large-scale public activism and social movements.

The European political system is composed of networks of agents (i.e., activists, decision-makers and organizations) with a range of social and political ties and is structured in nested and overlapping subsystems (i.e., national group, transnational political coalitions). Viewed through the lens of social tipping, European political dynamics present a 'social system', embedded within the broader international political and climate change governance community 'environment'. Driven by the *FridaysForFuture* movement (*16*) (among other things), a groundswell of bottom-up support for more proactive climate policies has recently developed among European citizens, resulting in routine mass demonstrations and historical wins for Green parties in the 2019 European Parliamentary Elections, as well as in federal elections in Austria, Belgium and Switzerland. The European political system could be moving towards a critical 'state', creating the conditions for a tipping process towards radical policy change, bringing European climate policy in line with the Paris Agreement. Accordingly, the European political system could constitute a potential 'social tipping element', where as it nears critical conditions, a small change to the system or its broader environment could lead to large-scale macroscopic changes, affected by cascading network dynamics and positive feedback mechanisms. Such transformations could involve establishing more aggressive mitigation strategies that connect goals (such as remaining below +2°C, 50% emissions reductions by 2030, zero carbon emissions by 2050) with measures and pathways that have a reasonable chance to achieve them (i.e., investment in negative emission technologies, increased carbon taxation policies etc.).

The *FridaysForFuture* movement has been pushing the European political system towards criticality, where it becomes more likely that the system will be propelled into a qualitatively different state. The movement was set off and inspired by a single Swedish high school student choosing to protest on the steps of the Riksdag for meaningful climate action. Greta Thunberg's protest quickly spread through the European social-political networks until more than a million students have been participating in weekly protests. This growing bottom-up pressure on the European climate policy-makers (*16*, *110*) has created an opening for significant policy change.

The European political system consists of embedded subsystems at multiple scales. At the national scale, for example, the German socio-political system responded strongly to the activities of the *FridaysForFuture* movement. Polling throughout 2019 in Germany suggested that the environment was the most important public policy challenge, ahead of other issues, such as the migration and financial crises. Drawing upon survey data collected monthly by the Politbarometer, 40–60% of Germans responded that the environment was an important problem in the Fall of 2019, a rapid increase from roughly 5% in the Fall of 2018 (Figure 3, Panels A and B). Since 2000, rarely more than 10% of Germans have viewed the environment as an important problem – a time period which includes the emergence of other large environmental movements in Germany, such as protests against nuclear energy in response to Fukushima. The specific upward shift in Germans viewing the environment as an important problem appears to coincide with the large-scale protests organized by *FridaysForFuture* in March, May and September of 2019.

Similarly, several national Western European Green Parties received historically strong electoral support in the May 2019 European Parliamentary Elections (such as in Belgium, Germany, Finland, France and Luxembourg). This increased support is also reflected in polling data in Germany, where the Green Party has been effectively equal with the conservative party as the preferred political party of German voters in the latter half of 2019 (Figure 3, Panels C and D). Subsequently, Germany introduced its first ever federal climate change laws, mandating that the country meet its 2030 goals (a ~55% reduction in GHG emissions) and establishing pathways to carbon neutrality by 2050. Currently, only a limited set of countries have enacted national climate change laws, and Germany is one of the largest and most diverse economies to propose such actions. This presents the possibility for policy diffusion and transfer to other states (*111*), particularly considering the influential role Germany plays within the European Union. Climate policy entrepreneurs could build upon momentum to further capitalize on windows of opportunity, pushing climate change proposals prominently into national and supra-national governmental agendas before the ephemeral moment passes (*112*).

The 2020 COVID-19 pandemic has placed new priorities on the policy agenda, also reflected in issue salience of climate change (see also Fig. S1 in Supplementary Materials). As political and behavioral responses to COVID-19 have led already to a significant temporary reduction in greenhouse gas emissions (*113*), this shock could be further leveraged to reinforce climate action –

future economic recovery packages should set European economies on a pathway towards carbon-neutrality, rather than return to the old normal (*114*, *115*). Drawing from this social tipping framework, the European political system may remain near a critical state. It remains unclear whether the COVID-19 shock has supplanted climate change, or whether both remain on the political agenda. For example, discussions of a "Green New Deal" remain at the core of COVID-19 economic recovery plans within the European Union.

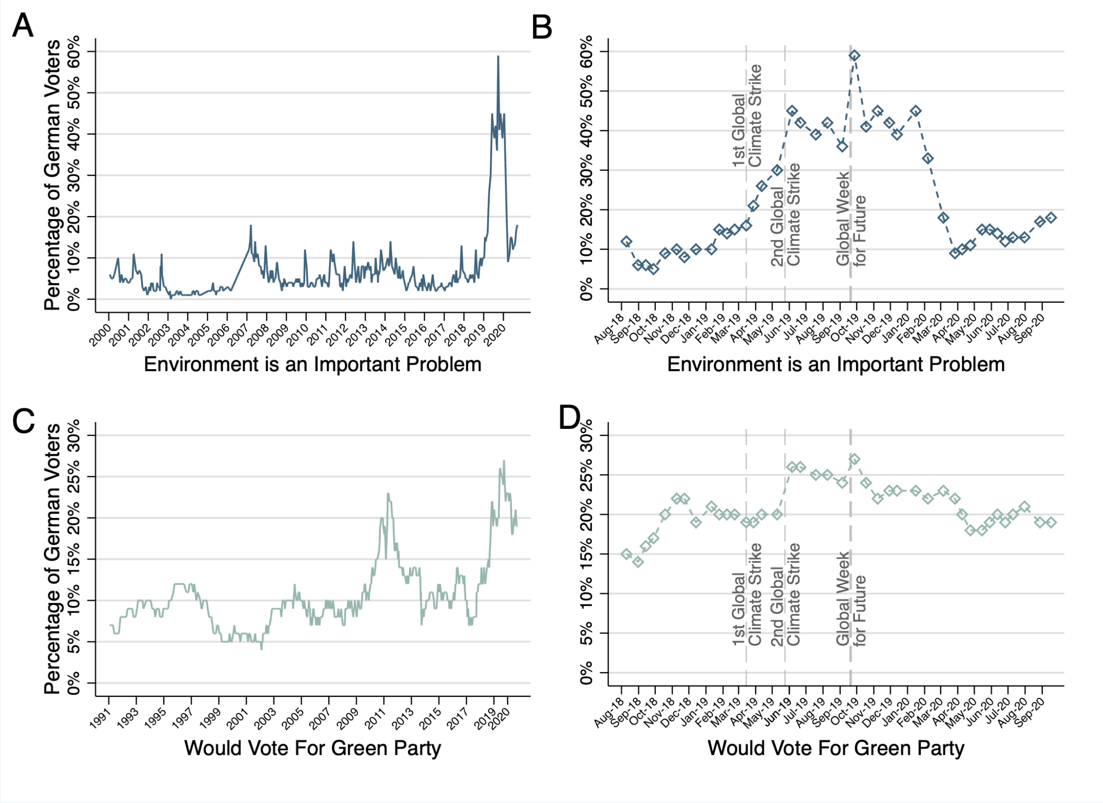


Figure 3: Environment as an issue and willingness to vote for the Green Party in Germany. *Percentages of potential German voters that list the environment as an important issue for the country and willingness to vote for the Green Party (Bündnis 90/Die Grünen) if the election were to be held "today". Panels (A) and (C) present monthly survey data from 2000 to September 2020 Panels (B) and (D) display monthly surveys from August 2018 – September 2020, showing the change since the beginning of Greta Thunberg's protest actions. Dotted grey vertical lines display days of global strikes organized by FridaysForFuture in March, May and September 2019. Data is collected by Forschungsgruppe Wahlen: Politbarometer .*

Implications for criticality

The sociopolitical dynamics have likely moved the Germany political subsystem further towards criticality, but it remains largely unknown whether this will result in tipping towards a qualitatively different state, in Germany or in the broader European political system. Rather, these judgements can *likely only be made in hindsight*, observing whether the system remained stable, moved towards criticality or experienced tipping dynamics. Such an analysis in line with the proposed framework requires specific process tracing, identifying the key moments, actors, networks, mechanisms affecting criticality, the triggering event (threshold), and the positive feedback dynamics propelling the system towards qualitative changes. Much attention is often paid to the specific triggering event, but it is rarely one single actor or action which accounts for the entirety of the tipping process.

Rather a full account needs to be made of all of the previous and related processes that have further placed the system towards criticality, allowing for such changes to become more likely. Accordingly, for a tipping process to occur at the scale of the entire European political system, moving it into a state of decarbonization that is aligned with the Paris Agreement, a series of additional social movements and protests, or other shifts within the system or the environment, may be required.

While we identify the role of *FridaysForFuture* in creating critical conditions, or potentially triggering the social transformations required for global sustainability, recent literature has identified further tipping candidates which could have generally "positive" effects on global sustainability. For example, divestment and reinvestment present candidates for rapid decarbonization and processes to achieve climate targets (*9*, *12*). In this case, intervention times range from years to decades, depending on the social structure level (*9*). Previous studies note that the adoption of technologies and behaviors such as rapid uptake of autonomously driven electric vehicles (if socially licensed), rapid change in dietary preferences reducing meat consumption and associated land-use and climate impacts can follow an epidemic-type model of diffusing across social networks (*13*, *15*).

Alternatively, social tipping processes can lead to states of criticality with less desirable outcomes: Recently it has been shown that climate change has contributed to the emergence of infections carried by mosquitoes, like dengue fever or Zika, which could be accelerated further by increased mobility, e.g., through denser air traffic networks (*75*). The thermal minimum for transmission of the Zika virus could in fact give rise to a threshold behaviour (*116*). Changes to the local environment may enact "push" factors, resulting in large scale migrations (*117*, *118*). Further, increased global mean temperature has been suggested to increase the likelihood of civil conflicts (*84*).

These social tipping processes are of great interest to policy makers, as it is desirable to potentially trigger or facilitate "positive" tipping (*11*, *13*), while at the same time, mitigating the effects of potential "negative" outcomes.

5. Discussion

Social tipping processes have been recently recognized as potentially key pathways for generating the necessary shifts for sustainability. Drawing upon this emerging field, this paper develops a framework for characterizing social tipping processes. We find that mechanisms underlying social tipping processes are more likely to exhibit the unique characteristics of agency, social-institutional and cultural network structures, they occur across different spatial and temporal scales to climate tipping, and the nature of tipping can be more complex. Social tipping processes thus present qualitatively different characteristics to those shared by climate tipping processes.

Accordingly, this paper develops a common framework for the unique characteristics of social tipping processes. We identify social tipping as a process, resultant of a complex system of drivers, resulting in shifting a system into a more (or less) critical state. It can thus serve to structure and inform future data analysis and process-based modelling exercises (*118*, *119*).

Even so, while there is an emerging focus on social tipping dynamics (*9*–*13*), there remains great difficulty in pinpointing tipping events and generalizing the emerging dynamics. Drawing from natural tipping dynamics**,** previous work on social tipping has often focused on identifying specific trigger events or critical thresholds in macroscopic system variables in analogy to identifying for instance critical temperature thresholds in the context of climate tipping (*10*). In natural systems the underlying dynamics are more deterministic and often can be directly observed, allowing for the identification of specific thresholds and events. While social systems comprise a much more open and complex system, one that is constantly adapting and where dynamics are often incredibly complex, interrelated and cannot be directly observed. Accordingly, one could observe the same event across ten similar social systems, and could potentially observe ten unique outcomes. As such, anticipating a specific trigger, making causal inferences, or having generalizability in expected effects are all greatly limited within social systems. Further, social tipping points are sometimes also understood as a point in time, rather than a point in a complex parameter space. Such an approach makes it difficult to identify social tipping processes, as they often do not contain easily observable macroscopic thresholds nor temporal markers for change.

Rather, a complex adaptive systems viewpoint is required, understanding the multitude of interrelated processes and social structures driving change, and not focusing on a single trigger or threshold. Accordingly, our framework proposed here focuses on identifying the processes and mechanisms of such change, and not a single triggering event, where the interplay of micro-level changes embedded within adaptive structural conditions can affect systemic changes.

The notion of a critical state is central within our framework. Changing conditions to the system's environment can cause it to enter more (or less) critical states, such that a single, or multiplicative action, can effect a systemic change. It is these changing conditions, and specifically the processes and dynamics underlying them, that are of analytical importance. Drawing upon the analogy of a tipping coal wagon (*15*), it is not the single, specific piece of coal that caused the wagon to tip, but rather the processes by which the wagon was filled with enough coal that any single piece (placed at a number of different locales) could cause such tipping. Accordingly, the specific triggering event of a social tipping process could be somewhat random or arbitrary, as the conditions are critical enough such that any event with enough magnitude could have triggered these dynamics.

It is therefore key to focus on the processes and mechanisms underlying the nature of such critical states which allow some trigger event to cause contagion dynamics. From social network models, we can deduce which kind of structural features make a system less resilient and thus more prone to social tipping (*119*). One example is polarization, where social network models and social media-based data analyses have shown that in polarized states with nearly disconnected network communities which in themselves are highly connected, contagion processes are more likely to occur (*120–122*). Behavioral experiments and corresponding conceptual modelling approaches suggest that minority groups can initiate social change dynamics in the emergence of new social conventions (*17*, *119*). Furthermore, a rich social science literature has noted an array of factors (i.e. political institutions, technological or behavioral adaptation, environmental, normative and attitudinal) effective in shifting the social conditions surrounding climate change (*14*). A better understanding of critical states as demanded by our framework may help to identify early warning signals that could possibly indicate that a social-ecological system is close to a critical state in specific situations (*30*, *123*).

Social tipping processes present a specific type of social change – characteristized by non-linear shifting states driving by positive feedbacks – which is similar to, but conceptually distinct from, other forms of social change. Similar to how we explore the differences between natural and social tipping processes, further research should engage with social tipping in comparison to other forms of social change (such as historical institutionalist perspectives, social movements, policy feedbacks, complex systems). One of the greatest challenges lies in dealing with multiple, entangled drivers of tipping processes on different scales – temporal, spatial or social structural levels – and different levels of agency and heterogeneous agents and subsystems. In order to further understand the dynamics arising from these various levels of agency, it is crucial to identify examples from different subfields (economics, political science, demographics). A key current limitation in applying our framework is finding and operationalizing empirical data describing actual spreading processes on networks across these different levels, particularly compared to macro-economic data and public opinion polls (*124*), even though first steps in this direction are being made (*125*, *126*). Particularly data on the social structures and networks is notoriously difficult to access. While there have been advances in developing modeling frameworks (*119*, *127*) to simulate social tipping dynamics, linking these theoretical modelling to empirical data and behavioral experiments requires more attention. Even if predictive modeling (i.e., the kind of deterministic, time-forward modeling we know from Earth System Models for instance) of such social dynamics in the sense of inferring time trajectories is very difficult or even conceptually unfeasible, such process-based modelling of social tipping dynamics can be very crucial to understand the nature of critical states also in real-world social situations. Lastly, we focus here specifically on social tipping processes relevant for mitigating climate change, or sustainability more broadly, fitting within the previous literature. But, such a framework for social tipping dynamics is generalizable to other areas of study and social phenomena (such as the 2020 rapid social movements and public opinion dynamics surrounding racial inequality in the United States).

While we explore one example of social tipping in detail, further inquiry is required to test the distinctiveness of social tipping processes, as well as the utility of the proposed definition to other social tipping processes. Systematizing the types of social tipping processes, and exemplary case studies, would help to further illustrate these forms of change. Research is also warranted into establishing typical timescales of social tipping; understanding how network structures affect social tipping dynamics; identifying typical network structures of systems entering critical states; discerning the temporal aspects of how effects travel through different social network structures; and gaining a better understanding of the origin of spreading processes. Data acquisition, analysis and process-based modelling could all play a role in this research agenda. A wealth of social media data is available to study potential social tipping processes. However, this kind of data has mostly yet to be adopted within the context of Earth System analysis and tipping dynamics.

Social tipping processes could be decisive for the future of the Earth System in the Anthropocene: some rapid shifts in social systems are, in fact, necessary to meet the targets of the Paris Agreement and the Sustainable Development Goals (*8*). While we focus here on processes relevant for future trajectories of the Earth system, we suggest that further analysis could use or adapt our definition to characterize other types of general social tipping processes (i.e. revolutions or rapid transformations). We also recognize that tipping processes within ecosystems present an interesting intermediary case between social and physical climate tipping as they typically incorporate characteristics from both realms. They are also crucial in determining future trajectories of the Earth system (see preliminary discussion in the SI). Understanding, identifying and potentially instigating some social tipping processes is highly relevant for the future of the Anthropocene, particularly with regard to the potential role in triggering rapid transformative change needed for effective Earth system stewardship (*9*, *11–13*).

References and Notes

1. T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, H. J. Schellnhuber, Tipping elements in the Earth's climate system. *Proc. Natl. Acad. Sci.* **105**, 1786–1793 (2008).
2. L. Favier, G. Durand, S. L. Cornford, G. H. Gudmundsson, O. Gagliardini, F. Gillet-Chaulet, T. Zwinger, A. J. Payne, A. M. Le Brocq, Retreat of Pine Island Glacier controlled by marine ice-sheet instability. *Nat. Clim. Change*. **4**, 117–121 (2014).
3. I. Joughin, B. E. Smith, B. Medley, Marine Ice Sheet Collapse Potentially Under Way for the Thwaites Glacier Basin, West Antarctica. *Science*. **344**, 735–738 (2014).
4. I. Joughin, R. B. Alley, Stability of the West Antarctic ice sheet in a warming world. *Nat. Geosci.* **4**, 506–513 (2011).
5. S. Drijfhout, S. Bathiany, C. Beaulieu, V. Brovkin, M. Claussen, C. Huntingford, M. Scheffer, G. Sgubin, D. Swingedouw, Catalogue of abrupt shifts in Intergovernmental Panel on Climate Change climate models. *Proc. Natl. Acad. Sci.* **112**, E5777–E5786 (2015).
6. K. Frieler, M. Meinshausen, A. Golly, M. Mengel, K. Lebek, S. D. Donner, O. Hoegh-Guldberg, Limiting global warming to 2 °C is unlikely to save most coral reefs. *Nat. Clim. Change*. **3**, 165–170 (2013).
7. T. P. Hughes, M. L. Barnes, D. R. Bellwood, J. E. Cinner, G. S. Cumming, J. B. Jackson, J. Kleypas, I. A. Van De Leemput, J. M. Lough, T. H. Morrison, others, Coral reefs in the Anthropocene. *Nature*. **546**, 82–90 (2017).
8. W. Steffen, J. Rockström, K. Richardson, T. M. Lenton, C. Folke, D. Liverman, C. P. Summerhayes, A. D. Barnosky, S. E. Cornell, M. Crucifix, J. F. Donges, I. Fetzer, S. J. Lade, M. Scheffer, R. Winkelmann, H. J.

Schellnhuber, Trajectories of the Earth System in the Anthropocene. *Proc. Natl. Acad. Sci.* **115**, 8252–8259 (2018).

10. M. Milkoreit, J. Hodbod, J. Baggio, K. Benessaiah, R. Calderón-Contreras, J. F. Donges, J.-D. Mathias, J. C. Rocha, M. Schoon, S. E. Werners, Defining tipping points for social-ecological systems scholarship—an interdisciplinary literature review. *Environ. Res. Lett.* **13**, 033005 (2018).
11. J. D. Tàbara, N. Frantzeskaki, K. Hölscher, S. Pedde, K. Kok, F. Lamperti, J. H. Christensen, J. Jäger, P. Berry, Positive tipping points in a rapidly warming world. *Curr. Opin. Environ. Sustain.* **31**, 120–129 (2018).
12. J. D. Farmer, C. Hepburn, M. C. Ives, T. Hale, T. Wetzer, P. Mealy, R. Rafaty, S. Srivastav, R. Way, Sensitive intervention points in the post-carbon transition. *Science*. **364**, 132–134 (2019).
13. T. M. Lenton, Tipping positive change. *Philos. Trans. R. Soc. B Biol. Sci.* **375**, 20190123 (2020).
14. K. Nyborg, J. M. Anderies, A. Dannenberg, T. Lindahl, C. Schill, M. Schlüter, W. N. Adger, K. J. Arrow, S. Barrett, S. Carpenter, others, Social norms as solutions. *Science*. **354**, 42–43 (2016).
15. R. E. Kopp, R. L. Shwom, G. Wagner, J. Yuan, Tipping elements and climate--economic shocks: Pathways toward integrated assessment. *Earths Future*. **4**, 346–372 (2016).
16. G. Hagedorn, P. Kalmus, M. Mann, S. Vicca, J. V. den Berge, J.-P. van Ypersele, D. Bourg, J. Rotmans, R. Kaaronen, S. Rahmstorf, H. Kromp-Kolb, G. Kirchengast, R. Knutti, S. I. Seneviratne, P. Thalmann, R. Cretney, A. Green, K. Anderson, M. Hedberg, D. Nilsson, A. Kuttner, K. Hayhoe, Concerns of young protesters are justified. *Science*. **364**, 139–140 (2019).
17. D. Centola, J. Becker, D. Brackbill, A. Baronchelli, Experimental evidence for tipping points in social convention. *Science*. **360**, 1116–1119 (2018).
18. J. Hoadley, A tilting water meter for purposes of experiment. *J. Frankl. Inst.* **117**, 273–278 (1884).
19. H. Poincaré, Sur l'équilibre d'une masse fluide animée d'un mouvement de rotation. *Acta Math.* **7**, 259–380 (1885).
20. T. C. Schelling, Dynamic models of segregation. *J. Math. Sociol.* **1**, 143–186 (1971).
21. M. Grodzins, Metropolitan segregation. *Sci. Am.* **197**, 33–41 (1957).
22. M. Granovetter, Threshold models of collective behavior. *Am. J. Sociol.* **83**, 1420–1443 (1978).
23. T. Kuran, Sparks and prairie fires: A theory of unanticipated political revolution. *Public Choice*. **61**, 41–74 (1989).
24. S. J. Gould, N. Eldredge, Punctuated equilibrium comes of age. *Nature*. **366**, 223 (1993).
25. M. Gladwell, *The tipping point: How little things can make a big difference* (Little, Brown and Company, Boston, MA, 2000).
26. T. Homer-Dixon, B. Walker, R. Biggs, A.-S. Crepin, C. Folke, E. F. Lambin, G. Peterson, J. Rockstrom, M. Scheffer, W. Steffen, M. Troell, Synchronous failure: the emerging causal architecture of global crisis. *Ecol. Soc.* (2015), doi:10.5751/ES-07681-200306.
27. S. Battiston, J. D. Farmer, A. Flache, D. Garlaschelli, A. G. Haldane, H. Heesterbeek, C. Hommes, C. Jaeger, R. May, M. Scheffer, Complexity theory and financial regulation. *Science*. **351**, 818–819 (2016).
28. K. Nyborg, M. Rege, On social norms: the evolution of considerate smoking behavior. *J. Econ. Behav. Organ.* **52**, 323–340 (2003).
29. J. Rockström, O. Gaffney, J. Rogelj, M. Meinshausen, N. Nakicenovic, H. J. Schellnhuber, A roadmap for rapid decarbonization. *Science*. **355**, 1269–1271 (2017).
30. M. Scheffer, *Critical transitions in nature and society* (Princeton University Press, 2009), vol. 16.
31. C. Folke, S. Carpenter, B. Walker, M. Scheffer, T. Elmqvist, L. Gunderson, C. S. Holling, Regime Shifts, Resilience, and Biodiversity in Ecosystem Management. *Annu. Rev. Ecol. Evol. Syst.* **35**, 557–581 (2004).
32. B. Walker, J. Meyers, Thresholds in Ecological and Social–Ecological Systems: a Developing Database. *Ecol. Soc.* **9** (2004), doi:10.5751/ES-00664-090203.
33. J. Tainter, *The collapse of complex societies* (Cambridge university press, 1990).
34. K. W. Butzer, Collapse, environment, and society. *Proc. Natl. Acad. Sci.* **109**, 3632–3639 (2012).
35. M. A. Janssen, T. A. Kohler, M. Scheffer, Sunk-cost effects and vulnerability to collapse in ancient societies. *Curr. Anthropol.* **44**, 722–728 (2003).
36. G. S. Cumming, G. D. Peterson, Unifying research on social--ecological resilience and collapse. *Trends Ecol. Evol.* **32**, 695–713 (2017).
37. J. C. Rocha, G. Peterson, Ö. Bodin, S. Levin, Cascading regime shifts within and across scales. *Science*. **362**, 1379–1383 (2018).
38. H. J. Schellnhuber, Tipping elements in the Earth System. *Proc. Natl. Acad. Sci.* **106**, 20561–20563 (2009).
39. A. Levermann, J. L. Bamber, S. Drijfhout, A. Ganopolski, W. Haeberli, N. R. Harris, M. Huss, K. Krüger, T. M. Lenton, R. W. Lindsay, others, Potential climatic transitions with profound impact on Europe. *Clim. Change*. **110**, 845–878 (2012).
40. C. A. Nobre, L. D. S. Borma, 'Tipping points' for the Amazon forest. *Curr. Opin. Environ. Sustain.* **1**, 28–36 (2009).
41. C. A. Nobre, G. Sampaio, L. S. Borma, J. C. Castilla-Rubio, J. S. Silva, M. Cardoso, Land-use and climate change risks in the Amazon and the need of a novel sustainable development paradigm. *Proc. Natl. Acad. Sci.*

113, 10759–10768 (2016).

42. M. Hirota, M. Holmgren, E. H. Van Nes, M. Scheffer, Global resilience of tropical forest and savanna to critical transitions. *Science*. **334**, 232–235 (2011).
43. A. Robinson, R. Calov, A. Ganopolski, Multistability and critical thresholds of the Greenland ice sheet. *Nat. Clim. Change*. **2**, 429 (2012).
44. S. Rahmstorf, Ocean circulation and climate during the past 120,000 years. *Nature*. **419**, 207 (2002).
45. L. Caesar, S. Rahmstorf, A. Robinson, G. Feulner, V. Saba, Observed fingerprint of a weakening Atlantic Ocean overturning circulation. *Nature*. **556**, 191 (2018).
46. Y. Cai, T. M. Lenton, T. S. Lontzek, Risk of multiple interacting tipping points should encourage rapid CO 2 emission reduction. *Nat. Clim. Change*. **6**, 520 (2016).
47. F. Westley, P. Olsson, C. Folke, T. Homer-Dixon, H. Vredenburg, D. Loorbach, J. Thompson, M. Nilsson, E. Lambin, J. Sendzimir, others, Tipping toward sustainability: emerging pathways of transformation. *Ambio*. **40**, 762 (2011).
48. M.-L. Moore, O. Tjornbo, E. Enfors, C. Knapp, J. Hodbod, J. A. Baggio, A. Norström, P. Olsson, D. Biggs, Studying the complexity of change: toward an analytical framework for understanding deliberate social-ecological transformations. *Ecol. Soc.* **19** (2014) (available at https://www.jstor.org/stable/26269689).
49. G. Feola, Societal transformation in response to global environmental change: A review of emerging concepts. *Ambio*. **44**, 376–390 (2015).
50. B. Ayyub, *Elicitation of Expert Opinions for Uncertainty and Risks* (CRC Press, Boca Raton, Florida, 2001).
51. L. Nash, The Agency of Nature or the Nature of Agency? *Environ. Hist.* **10**, 67–69 (2005).
52. B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory* (Oxford University Press, Oxford, UK, 2005).
53. European Environmental Agency, “Late lessons from early warnings:the precautionary principle 1896–2000,” *Environmental Issue Report* (22, European Environmental Agency, Copenhagen, 2001).
54. European Environmental Agency, “Late lessons from early warnings: Science, precaution, innovation,” *EEA Report* (1/2013, European Environmental Agency, Copenhagen, 2013).
55. A. Bandura, Toward a Psychology of Human Agency. *Perspect. Psychol. Sci.* **1**, 164–180 (2006).
56. A. Bandura, Human agency in social cognitive theory. *Am. Psychol.* **44**, 1175 (1989).
57. T. M. Lenton, B. Latour, Gaia 2.0. *Science*. **361**, 1066–1068 (2018).
58. E. Boyd, B. Nykvist, S. Borgström, I. A. Stacewicz, Anticipatory governance for social-ecological resilience. *AMBIO*. **44**, 149–161 (2015).
59. A. Hebinck, J. Vervoort, P. Hebinck, L. Rutting, F. Galli, Imagining transformative futures: participatory foresight for food systems change. *Ecol. Soc.* **23** (2018), doi:10.5751/ES-10054-230216.
60. L. Pereira, T. Hichert, M. Hamann, R. Preiser, R. Biggs, *Ecol. Soc.*, in press, doi:10.5751/ES-09907-230119.
61. T. Urdan, F. Pajares, *Selfefficacy beliefs of adolescents* (IAP, 2006).
62. M. Milkoreit, *Mindmade politics: The cognitive roots of international climate governance* (MIT Press, 2017).
63. M. Milkoreit, in *Reimagining Climate Change*, P. Wapner, H. Elver, Eds. (Routledge, New York, 2016), pp. 171–191.
64. A. Giddens, *The consequences of modernity* (Stanford University Press, Stanford, CA, 1990).
65. P. Bourdieu, L. J. Wacquant, *An invitation to reflexive sociology* (University of Chicago press, 1992).
66. R. A. Watson, E. Szathmáry, How Can Evolution Learn? *Trends Ecol. Evol.* **31**, 147–157 (2016).
67. C. Knappett, L. Malafouris, *Material Agency: Towards a Non-Anthropocentric Approach* (Springer, New York, 2008).
68. L. A. Brown, W. H. Walker, Prologue: Archaeology, Animism and Non-Human Agents. *J. Archaeol. Method Theory*. **15**, 297–299 (2008).
69. M. Newman, *Networks* (Oxford University Press, Oxford, UK, 2nd Edition., 2018).
70. E. Jablonka, M. Lamb, *Inheritance Systems and the Extended Synthesis* (Cambridge University Press, New York, 2020).
71. C. Castellano, S. Fortunato, V. Loreto, Statistical physics of social dynamics. *Rev. Mod. Phys.* **81**, 591–646 (2009).
72. M. Castells, G. Cardoso, others, *The network society: From knowledge to policy* (Johns Hopkins Center for Transatlantic Relations Washington, DC, 2006).
73. A. Giddens, *Runaway world: How globalization is reshaping our lives* (Taylor & Francis, 2003).
74. D. Harvey, *The condition of postmodernity* (Blackwell Oxford, 1989), vol. 14.
75. D. Brockmann, D. Helbing, The Hidden Geometry of Complex, Network-Driven Contagion Phenomena. *Science*. **342**, 1337–1342 (2013).
76. E. Ostrom, Polycentric systems for coping with collective action and global environmental change. *Glob. Environ. Change*. **20**, 550–557 (2010).
77. F. Kraas, C. Leggewie, P. Lemke, E. Matthies, D. Messner, N. Nakicenovic, H. J. Schellnhuber, S. Schlacke, U. Schneidewind, C. Brandi, C. Butsch, S. Busch, F. Hanusch, R. Haum, M. Jaeger-Erben, M. Köster, M. Kroll, C. Loose, A. Ley, D. Martens, I. Paulini, B. Pilardeaux, T. Schlüter, G. Schöneberg, A. Schulz, A.

Schwachula, B. Soete, B. Stephan, J. Sutter, K. Vinke, M. Wanner, *Humanity on the move: Unlocking the transformative power of cities* (WBGU - German Advisory Council on Global Change, Berlin, 2016; http://www.wbgu.de/en/flagship-reports/fr-2016-urbanization/).

78. J. Jankowski, R. Michalski, P. Bródka, A multilayer network dataset of interaction and influence spreading in a virtual world. *Sci. Data*. **4**, 170144 (2017).
79. D. M. Kahan, 'Ordinary science intelligence': a science-comprehension measure for study of risk and science communication, with notes on evolution and climate change. *J. Risk Res.* **20**, 995–1016 (2017).
80. W. D. Nordhaus, The Political Business Cycle. *Rev. Econ. Stud.* **42**, 169–190 (1975).
81. E. Dubois, Political business cycles 40 years after Nordhaus. *Public Choice*. **166**, 235–259 (2016).
82. A. Alesina, G. D. Cohen, N. Roubini, Electoral business cycle in industrial democracies. *Eur. J. Polit. Econ.* **9**, 1–23 (1993).
83. H. J. Schellnhuber, S. Rahmstorf, R. Winkelmann, Why the right climate target was agreed in Paris. *Nat. Clim. Change*. **6**, 649 (2016).
84. S. M. Hsiang, M. Burke, E. Miguel, Quantifying the Influence of Climate on Human Conflict. *Science*. **341** (2013), doi:10.1126/science.1235367.
85. S. Levin, T. Xepapadeas, A.-S. Crépin, J. Norberg, A. de Zeeuw, C. Folke, T. Hughes, K. Arrow, S. Barrett, G. Daily, P. Ehrlich, N. Kautsky, K.-G. Mäler, S. Polasky, M. Troell, J. R. Vincent, B. Walker, Social-ecological systems as complex adaptive systems: modeling and policy implications. *Environ. Dev. Econ.* **18**, 111–132 (2013).
86. C. S. Holling, Understanding the Complexity of Economic, Ecological, and Social Systems. *Ecosystems*. **4**, 390–405 (2001).
87. J. Miller, S. Page, *Adaptive Systems: An Introduction to Computational Models of Social Life* (Princeton University Press, Princeton, NJ, 2007).
88. C. Song, S. Havlin, H. A. Makse, Self-similarity of complex networks. *Nature*. **433**, 392–395 (2005).
89. C. Song, S. Havlin, H. A. Makse, Origins of fractality in the growth of complex networks. *Nat. Phys.* **2**, 275–281 (2006).
90. S. Thurner, R. Hanel, P. Klimek, *Introduction to the theory of complex system* (Oxford University Press, New York, 2018).
91. J.-D. Mathias, J. M. Anderies, J. Baggio, J. Hodbod, S. Huet, M. A. Janssen, M. Milkoreit, M. Schoon, Exploring non-linear transition pathways in social-ecological systems. *Sci. Rep.* **10**, 1–12 (2020).
92. J. Rockström, W. Steffen, K. Noone, Å. Persson, F. S. Chapin III, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, others, A safe operating space for humanity. *nature*. **461**, 472 (2009).
93. W. Steffen, K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R. Carpenter, W. de Vries, C. A. de Wit, C. Folke, D. Gerten, J. Heinke, G. M. Mace, L. M. Persson, V. Ramanathan, B. Reyers, S. Sörlin, Planetary boundaries: Guiding human development on a changing planet. *Science*. **347**, 1259855 (2015).
94. W. C. Clark, A. G. Harley, "Sustainability Science: Towards a Synthesis," *Sustainability Science Program* (Working Paper 2019-01, John F. Kennedy School of Government, Harvard University, Cambridge, MA).
95. J. Patterson, K. Schulz, J. Vervoort, S. van der Hel, O. Widerberg, C. Adler, M. Hurlbert, K. Anderton, M. Sethi, A. Barau, Exploring the governance and politics of transformations towards sustainability. *Environ. Innov. Soc. Transit.* **24**, 1–16 (2017).
96. F. W. Geels, Ontologies, socio-technical transitions (to sustainability), and the multi-level perspective. *Res. Policy*. **39**, 495–510 (2010).
97. F. W. Geels, The multi-level perspective on sustainability transitions: Responses to seven criticisms. *Environ. Innov. Soc. Transit.* **1**, 24–40 (2011).
98. M. Burke, S. M. Hsiang, E. Miguel, Climate and Conflict. *Annu. Rev. Econ.* **7**, 577–617 (2015).
99. B. Tonn, A. Hemrick, F. Conrad, Cognitive representations of the future: Survey results. *Futures*. **38**, 810–829 (2006).
100. C. Figueres, H. J. Schellnhuber, G. Whiteman, J. Rockström, A. Hobley, S. Rahmstorf, Three years to safeguard our climate. *Nat. News*. **546**, 593 (2017).
101. J. Rogelj, M. den Elzen, N. Höhne, T. Fransen, H. Fekete, H. Winkler, R. Schaeffer, F. Sha, K. Riahi, M. Meinshausen, Paris Agreement climate proposals need a boost to keep warming well below 2 °C. *Nature*. **534**, 631–639 (2016).
102. T. Rayner, A. Jordan, in *Climate Change Policy in the European Union: Confronting the Dilemmas of Mitigation and Adaptation?* (Cambridge University Press, Cambridge, UK, 2010), pp. 145–166.
103. C. F. Parker, C. Karlsson, M. Hjerpe, Assessing the European Union's global climate change leadership: from Copenhagen to the Paris Agreement. *J. Eur. Integr.* **39**, 239–252 (2017).
104. O. Geden, The Paris Agreement and the inherent inconsistency of climate policymaking. *Wiley Interdiscip. Rev. Clim. Change*. **7**, 790–797 (2016).
105. C. Wlezien, The Public as Thermostat: Dynamics of Preferences for Spending. *Am. J. Polit. Sci.* **39**, 981–1000 (1995).

106. S. N. Soroka, C. Wlezien, *Degrees of Democracy: Politics, Public Opinion and Policy* (Cambridge University Press, New York, 2010).
107. F. R. Baumgartner, B. D. Jones, *Agendas and instability in American politics* (University of Chicago Press, Chicago, 2010).
108. P. A. Sabatier, An advocacy coalition framework of policy change and the role of policy-oriented learning therein. *Policy Sci.* **21**, 129–168 (1988).
109. C. Weible, P. A. Sabatier, *Theories of the Policy Process* (Westview Press, New York, ed. 4th, 2017).
110. D. Evensen, The rhetorical limitations of the #FridaysForFuture movement. *Nat. Clim. Change*. **9**, 428–430 (2019).
111. C. R. Shipan, C. Volden, The Mechanisms of Policy Diffusion. *Am. J. Polit. Sci.* **52**, 840–857 (2008).
112. J. W. Kingdon, *Agendas, alternatives, and public policies* (HarperCollins College Publishers, New York, 1995).
113. C. Le Quéré, R. B. Jackson, M. W. Jones, A. J. P. Smith, S. Abernethy, R. M. Andrew, A. J. De-Gol, D. R. Willis, Y. Shan, J. G. Canadell, P. Friedlingstein, F. Creutzig, G. P. Peters, Temporary reduction in daily global CO 2 emissions during the COVID-19 forced confinement. *Nat. Clim. Change*. **10**, 647–653 (2020).
114. R. Hanna, Y. Xu, D. G. Victor, After COVID-19, green investment must deliver jobs to get political traction. *Nature*. **582**, 178–180 (2020).
115. D. Rosenbloom, J. Markard, A COVID-19 recovery for climate. *Science*. **368**, 447–447 (2020).
116. B. Tesla, L. R. Demakovsky, E. A. Mordecai, S. J. Ryan, M. H. Bonds, C. N. Ngonghala, M. A. Brindley, C. C. Murdock, Temperature drives Zika virus transmission: evidence from empirical and mathematical models. *Proc. R. Soc. B Biol. Sci.* **285**, 20180795 (2018).
117. R. McLeman, B. Smit, Migration as an Adaptation to Climate Change. *Clim. Change*. **76**, 31–53 (2006).
118. R. Jennissen, Causality Chains in the International Migration Systems Approach. *Popul. Res. Policy Rev.* **26**, 411–436 (2007).
119. M. Wiedermann, E. K. Smith, J. Heitzig, J. F. Donges, A network-based microfoundation of Granovetter's threshold model for social tipping. *Sci. Rep.* **10**, 11202 (2020).
120. P. Törnberg, Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLOS ONE*. **13**, e0203958 (2018).
121. V. V. Vasconcelos, S. A. Levin, F. L. Pinheiro, Consensus and polarization in competing complex contagion processes. *J. R. Soc. Interface*. **16**, 20190196 (2019).
122. M. Del Vicario, G. Vivaldo, A. Bessi, F. Zollo, A. Scala, G. Caldarelli, W. Quattrociocchi, Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Sci. Rep.* **6**, 37825 (2016).
123. C. T. Bauch, R. Sigdel, J. Pharaon, M. Anand, Early warning signals of regime shifts in coupled human–environment systems. *Proc. Natl. Acad. Sci.* **113**, 14560–14567 (2016).
124. D. Helbing, S. Bishop, R. Conte, P. Lukowicz, J. B. McCarthy, FuturICT: Participatory computing to understand and manage our complex world in a more sustainable and resilient way. *Eur. Phys. J. Spec. Top.* **214**, 11–39 (2012).
125. V. Sekara, A. Stopczynski, S. Lehmann, Fundamental structures of dynamic social networks. *Proc. Natl. Acad. Sci.* **113**, 9977–9982 (2016).
126. P. Sapiezynski, A. Stopczynski, D. D. Lassen, S. Lehmann, Interaction data from the Copenhagen Networks Study. *Sci. Data*. **6**, 315 (2019).
127. J. F. Donges, J. Heitzig, W. Barfuss, M. Wiedermann, J. A. Kassel, T. Kittel, J. J. Kolb, T. Kolster, F. Müller-Hansen, I. M. Otto, K. B. Zimmerer, W. Lucht, Earth system modeling with endogenous and dynamic human societies: the copan:CORE open World–Earth modeling framework. *Earth Syst. Dyn.* **11**, 395–413 (2020).
128. B. Schuldt, A. Buras, M. Arend, Y. Vitasse, C. Beierkuhnlein, A. Damm, M. Gharun, T. E. E. Grams, M. Hauck, P. Hajek, H. Hartmann, E. Hiltbrunner, G. Hoch, M. Holloway-Phillips, C. Körner, E. Larysch, T. Lübbe, D. B. Nelson, A. Rammig, A. Rigling, L. Rose, N. K. Ruehr, K. Schumann, F. Weiser, C. Werner, T. Wohlgemuth, C. S. Zang, A. Kahmen, A first assessment of the impact of the extreme 2018 summer drought on Central European forests. *Basic Appl. Ecol.* **45**, 86–103 (2020).

Acknowledgments

General: We are very grateful to William C. Clark, Anne-Sophie Crépin, Niklas Harring, Matthew Ives, J. Doyne Farmer, Wolfgang Lucht, and John Schellnhuber, for providing helpful insights and comments. We thank the participants of two DominoES workshops on social tipping dynamics held at GESIS Leibniz Institute for Social Science, Cologne, in summer 2018 and 2019 for foundational and framing discussions.

Funding: This work has been supported by the Leibniz Association (project DominoES), the Stordalen Foundation via the Planetary Boundary Research Network (PB.net), the Earth League's EarthDoc programme, IRTG 1740 funded by DFG and FAPESP, and the European Research Council (ERC advanced grant project ERA: Earth resilience in the Anthropocene, ERC-2016-ADG-743080). T.M.L.'s contribution was supported by the Leverhulme Trust (RPG-2018-046).

Author Contributions: Drawing upon the concepts developed in the expert elicitation workshop, R.W., J.F.D., E.K.S., M.M. and T.M.L structured the conceptualization into the resultant framework and wrote the paper with the support of all co-authors. All co-authors contributed to the discussion of the manuscript. M.W. analyzed data and created Fig. 1. R.W. and E.K.S. created Fig. 2. E.K.S. analyzed data and created Fig. 3. J.H. derived the mathematical definition of social tipping processes (Sect. S1).

Competing Interests: The authors declare no competing interests.

Supplementary Materials

S1: A mathematical definition of social tipping processes

In this section, we give a more formal version of the definition of 'social tipping process' given in the main text, as a reference for mathematically inclined readers.

After defining what we mean by a social system and its environment, we first classify their possible states into critical, unmanageable, uncritical, and tippable conditions, and then finally define the notions of prevention time and triggering time.

By a *social system*, Σ, we mean a set of agents together with a network-like social structure, that interacts in some form with the rest of the world, called the *environment*, E, of the system, such that, if no "perturbation" or deliberate "influence" by some decision-maker occurs, Σ and E together can only follow certain "quasi-inertial" (or "default") trajectories restricted by the agency of the system's agents. Let $x_{(t)}$ and $y_{(t)}$ denote the *states* that Σ and E are actually in at time t.

A *critical condition* for the system is a pair of possible system and environment states, (x^*, y^*), such that there exists another possible pair of states, (x', y'), with the following properties:

1. The state pair (x', y') is no further away in state space from (x^*, y^*) than a certain "small" distance, ϵ, that represents the possible magnitude of "local" perturbations in Σ (affecting only few agents or network links directly) or small changes in E that are considered sufficiently "likely" to care about, with respect to some suitable distance function d. In other words, $d((x', y'), (x^*, y^*)) < \epsilon$.
2. If Σ and E were in state (x', y') at any time t', there is a quasi-inertial trajectory that would move Σ at some later time $t'' > t'$ into some state x'' that is "qualitatively" different from x^*. This move represents a "global" (i.e., affecting a very large fraction of the agents) and "significant" change in the system (but not necessarily in its environment).

If such a change actually happens, the time point t' (not the state!) at which it starts may be called the *tipping point* or less ambiguously the *triggering time point*, and the system behavior within the time interval from t' to t'' is called the corresponding *tipping process*. An *uncritical* condition for Σ and E then is any pair of states that is not critical.

A critical condition is *unmanageable* for an actor that may influence Σ or E in some way if there exists a possible pair of states, (x', y'), with $d((x', y'), (x^*, y^*)) < \epsilon$ and the following property:

- Assume that Σ and E were in state (x', y') at any time t' and afterwards the state of Σ and E would follow any trajectory $(x(t), y(t))_{t \geq t'}$ that the actor can force it to follow. Then the resulting trajectory would still move Σ at some time $t'' > t'$ into some state x'' (which will usually depend on the influence exerted) that is qualitatively different from x^*.

Similarly, an uncritical condition, (x°, y°), is *tippable* by a decision maker if there is a possible trajectory $(x(t), y(t))_{t \geq t'}$, starting in (x°, y°) at some time t', that the decision maker can force Σ and E to follow, and this trajectory would move Σ into some state x'' at some time $t'' > t'$ that is qualitatively different from x° (a tippable uncritical state roughly corresponds to what others call a 'sensitive intervention point').

At any time at which the system is not in an unmanageable critical state, the *prevention time* is the time interval it takes before some quasi-inertial trajectory has moved it into an unmanageable critical state. In other words, at time zero it is the largest time interval T so that, when no intervention takes place until time T, for all $t > 0$ with $t < T$, the system would not be in an unmanageable critical state at time t.

Similarly, at any time at which the system is in a tippable uncritical state, the *triggering time* is the time interval it takes before some quasi-inertial trajectory has moved it into an uncritical state that is no longer tippable. In other words, at time zero it is the largest time interval T so that, when no intervention takes place until time T, for all $t > 0$ with $t < T$, the system would not be in a tippable uncritical state at time t.

We only consider social tipping processes for which the prevention or triggering time is smaller than some *intervention time horizon*.

S2 Ecosystem tipping as intermediary case

Ecosystem tipping processes share properties of physical climate tipping dynamics in atmosphere, ocean and cryosphere in that they can often be described by a common driver, as well as that of deliberative social tipping elements in that they have adaptive capacity, and can therefore be regarded as intermediate. But, as previously noted, human agential capacity is far greater than those of other species.

Similarly to human social systems, ecosystems are comprised of interacting living organisms, they can be viewed as networks with components that can adapt (e.g., food webs). This is different from physical tipping elements such as the cryosphere elements (e.g., melting of permafrost) which do not typically exhibit the same networked structures. Within the nominally 'climate' tipping elements are some major biomes – notably boreal forests, the Amazon rainforest, and coral reefs – that are composed of living organisms and exhibit ecological network structures. Indeed changing interactions between the living elements of these systems may be key to tipping dynamics – for example epidemic bark beetle infestation of boreal forests triggered by climate warming allowing the beetles to complete two life cycles rather than one within a season (*128*). Thus these biotic tipping elements lie towards smaller scale ecosystems in the continuum, and tend to be more closely related to social systems in spatial and temporal scales compared to the typically much larger and more slowly changing physical climate tipping elements.

These differences give rise to a proposed ordering of tipping elements, ranging from (1) the physical climate tipping elements via (2) ecosystem tipping elements to (3) social tipping elements (Table S1).

Table S1: Proposed ordering of tipping processes ranging from physical climate tipping processes via ecosystem tipping processes to social tipping processes.

Properties	**Physical climate tipping processes**	**Ecological tipping processes**	**Social tipping processes**
Degree of agency	*Low/Absent*	*Intermediate*	*High*
Network structure	*Uncommon*	*Common*	*Common*
Temporal-spatial scales	*Slower and larger*	*Faster and smaller*	*Faster and smaller*
Degree of complexity	*Lower*	*Intermediate*	*High*

Figure S1:

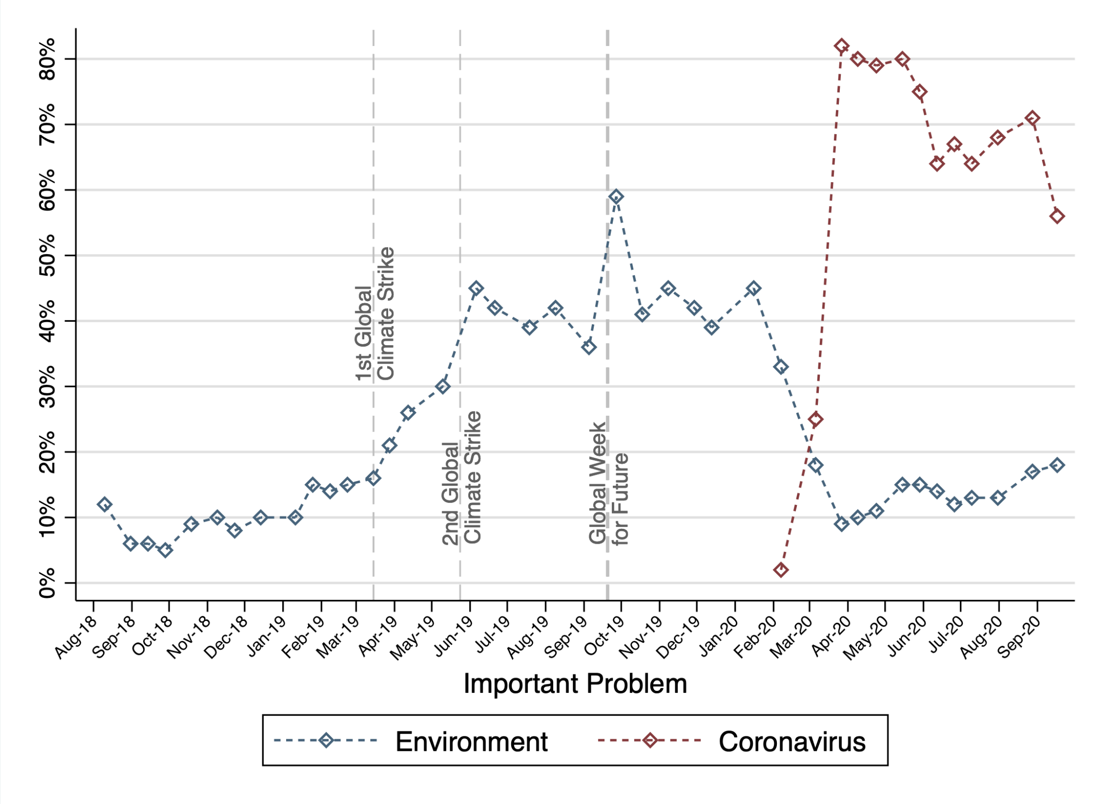


Figure S1: Environment and Corona as an important issue in Germany. *Percentages of potential German voters that list the environment and the Coronavirus as an important issue for the country*

from August 2018 – September 2020, showing the change since the beginning of Greta Thunberg's protest actions. Dotted grey vertical lines display days of global strikes organized by FridaysForFuture in March, May and September 2019. Data is collected by Forschungsgruppe Wahlen: Politbarometer .

Earth Syst. Dynam., 11, 395–413, 2020
https://doi.org/10.5194/esd-11-395-2020

Earth System Dynamics
Open Access
EGU

Earth system modeling with endogenous and dynamic human societies: the copan:CORE open World–Earth modeling framework

Jonathan F. Donges[1,2,*], **Jobst Heitzig**[1,*], **Wolfram Barfuss**[1,3], **Marc Wiedermann**[1], **Johannes A. Kassel**[1,4], **Tim Kittel**[3], **Jakob J. Kolb**[1,3], **Till Kolster**[1,3], **Finn Müller-Hansen**[1,5], **Ilona M. Otto**[1], **Kilian B. Zimmerer**[1,6], and **Wolfgang Lucht**[1,7,8]

[1]Earth System Analysis and Complexity Science, Potsdam Institute for Climate Impact Research,
Member of the Leibniz Association, Telegrafenberg A31, 14473 Potsdam, Germany
[2]Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden
[3]Department of Physics, Humboldt University, Newtonstr. 15, 12489 Berlin, Germany
[4]Max Planck Institute for the Physics of Complex Systems, Nöthnitzer Straße 38, 01187 Dresden, Germany
[5]Mercator Research Institute on Global Commons and Climate Change (MCC),
EUREF Campus 19, Torgauer Straße 12–15, 10829 Berlin, Germany
[6]Department of Physics and Astronomy, University of Heidelberg,
Im Neuenheimer Feld 226, 69120 Heidelberg, Germany
[7]Department of Geography, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany
[8]Integrative Research Institute on Transformations of Human-Environment Systems,
Humboldt University, Unter den Linden 6, 10099 Berlin, Germany
[*]The first two authors share the lead authorship.

Correspondence: Jonathan F. Donges (donges@pik-potsdam.de) and Jobst Heitzig (heitzig@pik-potsdam.de)

Received: 29 December 2017 – Discussion started: 15 January 2018
Revised: 3 March 2020 – Accepted: 7 March 2020 – Published: 4 May 2020

Abstract. Analysis of Earth system dynamics in the Anthropocene requires explicitly taking into account the increasing magnitude of processes operating in human societies, their cultures, economies and technosphere and their growing feedback entanglement with those in the physical, chemical and biological systems of the planet. However, current state-of-the-art Earth system models do not represent dynamic human societies and their feedback interactions with the biogeophysical Earth system and macroeconomic integrated assessment models typically do so only with limited scope. This paper (i) proposes design principles for constructing world–Earth models (WEMs) for Earth system analysis of the Anthropocene, i.e., models of social (world)–ecological (Earth) coevolution on up to planetary scales, and (ii) presents the copan:CORE open simulation modeling framework for developing, composing and analyzing such WEMs based on the proposed principles. The framework provides a modular structure to flexibly construct and study WEMs. These can contain biophysical (e.g., carbon cycle dynamics), socio-metabolic or economic (e.g., economic growth or energy system changes), and sociocultural processes (e.g., voting on climate policies or changing social norms) and their feedback interactions, and they are based on elementary entity types, e.g., grid cells and social systems. Thereby, copan:CORE enables the epistemic flexibility needed for contributions towards Earth system analysis of the Anthropocene given the large diversity of competing theories and methodologies used for describing socio-metabolic or economic and socio-cultural processes in the Earth system by various fields and schools of thought. To illustrate the capabilities of the framework, we present an exemplary and highly stylized WEM implemented in copan:CORE that illustrates how endogenizing sociocultural processes and feedbacks such as voting on climate policies based on socially learned environmental awareness could fundamentally change macroscopic model outcomes.

Published by Copernicus Publications on behalf of the European Geosciences Union.

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

1 Introduction

In the Anthropocene, Earth system dynamics are equally governed by two kinds of internal processes: those operating in the physical, chemical and biological systems of the planet and those occurring in its human societies, their cultures and economies (Schellnhuber, 1998, 1999; Crutzen, 2002; Lucht and Pachauri, 2004; Steffen et al., 2018). The history of global change is the history of the increasing planetary-scale entanglement and strengthening of feedbacks between these two domains (Lenton and Watson, 2011). Therefore, Earth system analysis of the Anthropocene requires closing the loop by integrating the dynamics of complex human societies into integrated *whole* Earth system models (Verburg et al., 2016; Donges et al., 2017a, b; Calvin and Bond-Lamberty, 2018). Such models need to capture the coevolving dynamics of the social (the world of human societies) and natural (the biogeophysical Earth) spheres of the Earth system on up to global scales and are referred to as world–Earth models (WEMs) in this article. In pursuing this interdisciplinary integration effort, world–Earth modeling can benefit from and build upon the work done in fields such as social–ecological systems (Berkes et al., 2000; Folke, 2006) and coupled human and natural systems (Liu et al., 2007) research or large-scale behavioral land-use (Arneth et al., 2014; Rounsevell et al., 2014) and socio-hydrological modeling (Di Baldassarre et al., 2017). However, it emphasizes more the study of planetary-scale interactions between human societies and parts of the Earth's climate system such as atmosphere, ocean and the biosphere, instead of more local and regional-scale interactions with natural resources that these fields have typically focused on in the past (Donges et al., 2018).

The contribution of this paper is twofold: first, following a more detailed motivation (Sect. 1.1), general theoretical considerations and design principles for a novel class of integrated WEMs are discussed (Sect. 1.2) and WEMs are elaborated in the context of existing global modeling approaches (Sect. 1.3). Second, after a short overview of the copan:CORE open World–Earth modeling framework (Sect. 2), an exemplary full-loop WEM is presented and studied (Sect. 3), showing the relevance of internalizing sociocultural processes. Finally, Sect. 4 concludes the paper.

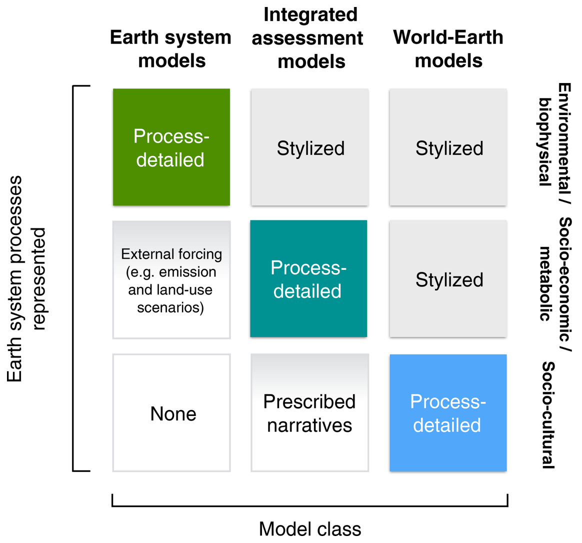
1.1 State of the art and research gaps in Earth system analysis

Computer simulation models are pivotal tools for gaining scientific understanding and providing policy advice for addressing global change challenges such as anthropogenic climate change or rapid degradation of biosphere integrity and their interactions (Rockström et al., 2009; Steffen et al., 2015). At present, two large modeling enterprises considering the larger Earth system in the Anthropocene are mature (van Vuuren et al., 2016). (i) Biophysical Earth system models (ESMs) derived from and built around a core of atmosphere–ocean general circulation models that are evaluated using storyline-based socioeconomic scenarios to study anthropogenic climate change and its impacts on human societies (e.g., representative concentration pathways, RCPs) (Stocker et al., 2013). (ii) Socioeconomic integrated assessment models (IAMs) are operated using storyline-based socioeconomic baseline scenarios (e.g., shared socioeconomic pathways, SSPs; Edenhofer et al., 2014) and evaluate technology and policy options for mitigation and adaption leading to different emission pathways. There is a growing number of intersections, couplings and exchanges between the biophysical and socioeconomic components of these two model classes for increasing their consistency (van Vuuren et al., 2012; Foley et al., 2016; Dermody et al., 2018; Robinson et al., 2018; Calvin and Bond-Lamberty, 2018).

However, the existing scientific assessment models of global change only include dynamic representations of the sociocultural dimensions of human societies to a limited degree – if at all (Fig. 1), i.e., the diverse political and economic actors, the factors influencing their decisions and behavior, their interdependencies constituting social network structures and institutions (Verburg et al., 2016; Donges et al., 2017a, b), and the broader technosphere they created (Haff, 2012, 2014). In IAMs, these sociocultural dimensions are partly represented by different socioeconomic scenarios (e.g., SSPs), providing the basis for different emission pathways. These are in turn used in ESMs as external forcing, constraints and boundary conditions to the modeled Earth system dynamics. However, a dynamic representation would be needed to explore how changes in the global environment influence these sociocultural factors and vice versa.

There are large differences in beliefs, norms, economic interests and political ideologies of various social groups and their metabolic profiles, which are related to their access and use of energy and resources (Fischer-Kowalski, 1997; Otto et al., 2019; Lenton et al., 2016; Lenton and Latour, 2018). Historical examples show that these differences might lead to rapid social changes, revolutions and sometimes also devastating conflicts, wars and collapse (Betts, 2017; Cumming and Peterson, 2017). In other cases, the inability to establish effective social institutions controlling resource access might lead to unsustainable resource use and resource degradation (see the discussion around the tragedy of the commons, Ostrom, 1990; Jager et al., 2000; Janssen, 2002). Climate change is a paradigmatic example of a global commons that needs global institutional arrangements for the use of the atmosphere as a deposit for greenhouse gas emissions if substantial environmental and social damage is to be avoided in the future (Edenhofer et al., 2015; Schellnhuber et al., 2016b; Otto et al., 2017).

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).



	Earth system models	Integrated assessment models	World-Earth models	
	Process-detailed	Stylized	Stylized	Environmental / biophysical
	External forcing (e.g. emission and land-use scenarios)	Process-detailed	Stylized	Socio-economic / metabolic
	None	Prescribed narratives	Process-detailed	Socio-cultural

Figure 1. World–Earth models (WEMs) in the space of model classes used for scientific analysis of global change. It is shown to what degree current Earth system models, integrated assessment models and WEMs cover environmental or biophysical, socioeconomic or metabolic, and sociocultural processes. The term "process-detailed" indicates the types of Earth system processes that the different model classes typically focus on representing. However, also in these core areas the level of detail may range from very stylized to complex and highly structured.

In order to explore the risks, dangers and opportunities for sustainable development, it is important to understand how biophysical, socioeconomic and sociocultural processes influence each other (Donges et al., 2018), how institutional and other social processes function, and which tipping elements can emerge from the interrelations of the subsystems (Lenton et al., 2008; Kriegler et al., 2009; Cai et al., 2016; Kopp et al., 2016; Otto et al., 2020a). To address these questions, the interactions of social systems and the natural Earth system can be regarded as part of a planetary social–ecological system (SES) or world–Earth system, extending the notion of SES beyond its common usage to describe systems on local scales (Berkes et al., 2000; Folke, 2006). This dynamical systems perspective allows us to explore under which preconditions the maintenance of planetary boundaries (Rockström et al., 2009; Steffen et al., 2015), i.e., a Holocene-like state of the natural Earth system, can be reconciled with human development to produce an ethically defensible trajectory of the whole Earth system (i.e., sustainable development) (Raworth, 2012; Steffen et al., 2018).

1.2 World–Earth modeling: contributions towards Earth system analysis of the Anthropocene

To this end, the case has been made that substantial efforts are required to advance the development of integrated world–Earth system models for the study of the Anthropocene (Verburg et al., 2016; Donges et al., 2017a, b; Calvin and Bond-Lamberty, 2018). The need for developing such next-generation social–ecological models has been recognized in several subdisciplines of global change science dealing with socio-hydrology (Di Baldassarre et al., 2017; Keys and Wang-Erlandsson, 2018), land-use dynamics (Arneth et al., 2014; Robinson et al., 2018) and the globalized food–water–climate nexus (Dermody et al., 2018). While in recent years there has been some progress in developing stylized models that combine sociocultural with economic and natural dynamics (e.g., Janssen and De Vries, 1998; Kellie-Smith and Cox, 2011; Garrett, 2014; Motesharrei et al., 2014; Wiedermann et al., 2015; Heck et al., 2016; Barfuss et al., 2017; Nitzbon et al., 2017; Beckage et al., 2018), more advanced and process-detailed WEMs are not yet available for studying the deeper past and the longer-term Anthropocene future of this coupled system. The research program investigating the dynamics and resilience of the world–Earth system in the Anthropocene can benefit from recent advances in the theory and modeling of complex adaptive systems (Farmer et al., 2015; Verburg et al., 2016; Donges et al., 2017a, b; Calvin and Bond-Lamberty, 2018). When advancing beyond stylized modeling, a key challenge for world–Earth modeling is the need to take into account the agency of heterogeneous social actors and global-scale adaptive networks carrying and connecting social, economic and ecological processes that shape social–ecological coevolution (Otto et al., 2020b).

A number of new developments make it attractive to revisit the challenge of building such WEMs now. Due to the huge progress in computing, comprehensive Earth system modeling is advancing fast. And with the ubiquity of computers and digital communication for simulation and data acquisition in daily life (Otto et al., 2015), efforts to model complex social systems are increased and become more concrete. Recent advances, for example in the theory of complex (adaptive) systems, computational social sciences, social simulation and social–ecological system modeling (Farmer and Foley, 2009; Farmer et al., 2015; Helbing et al., 2012; Müller-Hansen et al., 2017; Schill et al., 2019) make it feasible to include some important macroscopic dynamics of human societies regarding, among others, the formation of institutions, values and preferences and various processes of decision-making in a model of the whole Earth system, i.e., the physical Earth including its socially organized and mentally reflexive humans. Furthermore, new methodological approaches are developing fast that allow representing crucial aspects of social systems, such as adaptive complex networks (Gross and Blasius, 2008; Snijders et al., 2010). Finally, initiatives such as Future Earth (Fu-

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

ture Earth, 2014), the Earth League (Rockström et al., 2014, https://www.the-earth-league.org/, last access: 1 April 2020) and the Open Modeling Foundation (Barton and The Open Modeling Foundation, 2019) provide a basis for inter- and transdisciplinary research that could support such an ambitious modeling program.

It is important to emphasize that despite these advances, integrated world–Earth modeling studies still face challenges particularly in the areas of selecting and managing the appropriate level of model complexity, mathematical representations of human behavior and social dynamics, costs of computation and model development, and data availability and consistency, as highlighted by a recent literature review (Calvin and Bond-Lamberty, 2018). While at least a subset of these challenges tends to apply to many other ambitious modeling projects in diverse fields, they have been used as a basis of criticism of past human-environment modeling exercises such as the classic WORLD3 model in the "Limits to growth" studies (Meadows et al., 1972). To address these challenges, as we detail in Sect. 2, world–Earth system modeling should be developed following a modular approach, allowing for the intercomparison of a diversity of modeling approaches and corresponding extensive robustness and uncertainty analyses (Verburg et al., 2016). Model types and complexity levels should be selected carefully depending on the research questions of interest (van Vuuren et al., 2016). Community development is needed to foster the necessary interdisciplinary collaboration and to develop common protocols and ontologies for data, model simulations and intercomparison projects (Otto et al., 2015; Verburg et al., 2016; Calvin and Bond-Lamberty, 2018; Barton and The Open Modeling Foundation, 2019).

1.2.1 Research questions for world–Earth modeling

We envision world–Earth modeling to be complementary to existing simulation approaches for the analysis of global change. WEMs are not needed where the focus is on the study of the biophysical and climatic implications of certain prescribed socioeconomic development pathways (e.g., in terms of emission and land-use scenarios), since this is the domain of Earth system models as used in the World Climate Research Programme's Coupled Model Intercomparison Project (CMIP) (Eyring et al., 2016) that provides input to the Intergovernmental Panel on Climate Change (IPCC) reports. Similarly, WEMs are not the tool of choice if the interest is in the normative macroeconomic projection of optimal socioeconomic development and policy pathways internalizing certain aspects of climate dynamics, e.g., the analysis of first- or second-best climate change mitigation pathways, since this is the domain of state-of-the-art integrated assessment models.

In turn, WEMs as envisioned by us here are needed when the research questions at hand require the explicit and internalized representation of sociocultural processes and their feedback interactions with biophysical and socioeconomic dynamics in the Earth system. In the following, we give examples of research questions of this type that could be studied with WEMs in the future, as they have been already elaborated in more detail by, e.g., Verburg et al. (2016), Donges et al. (2017a, b) and Beckage et al. (2018):

1. What are the relative strengths of feedback interactions between biophysical processes in the climate system and processes of decision-making and behavioral change in human societies (Calvin and Bond-Lamberty, 2018)? For example, what is their influence on the uncertainty of projected global warming under different emission and land-use scenarios (Beckage et al., 2018)?

2. What are the sociocultural, socioeconomic and environmental preconditions for sustainable development towards and within a "safe and just" operating space for humankind (Barfuss et al., 2018; O'Neill et al., 2018), i.e., for a trajectory of the Earth system that eventually neither violates precautionary planetary boundaries (Rockström et al., 2009; Steffen et al., 2015) nor acceptable social foundations (Raworth, 2012)?

3. A more specific example of the previous questions is: how can major socioeconomic transitions towards a decarbonized social metabolism, such as a transformation of the food and agricultural systems towards a sustainable, reduced-meat diet that is in line with recent recommendations by the EAT-Lancet Commission on healthy diets (Willett et al., 2019), be brought about in view of the strong sociocultural drivers of current food-related and agricultural practices and the reality of the political economy in major food-producing countries? And how would their progress be influenced by realized or anticipated tipping of climatic tipping elements like the West Antarctic Ice Sheet (Wiedermann et al., 2019)?

4. Under which conditions can cascading interactions between climatic (e.g., continental ice sheets or major biomes such as the Amazon rain forest) and potential social tipping elements (e.g., in attitudes towards ongoing or anticipated climate change or eco-migration) be triggered and how can they be governed (Schellnhuber et al., 2016a; Steffen et al., 2018; Wiedermann et al., 2019)? What are implications for biophysical and social–ecological dimensions of Earth system resilience in the Anthropocene (Donges et al., 2017a)?

5. How do multilevel coalition formation processes (like the one modeled in Heitzig and Kornek (2018) assuming a static climate) interact with Earth system dynamics via changes in regional damage functions, mitigation costs, and realized or anticipated distributions of extreme events that drive changes in public opinions, which in turn influence the ratification of international

treaties and the implementation of domestic climate policies?

6. How do certain social innovations including technology, policies or behavioral practices diffuse in heterogeneous agent networks that could have global-scale impacts on planetary-boundary dimensions (e.g., Farmer et al., 2019; Tàbara et al., 2018; Otto et al., 2020a)? Which factors, such as network structure, information access as well as information feedback and update time, affect the innovation uptake? What are the impacts of a certain social innovation uptake on different agent groups (e.g., on agents with different economic, social or cultural endowment)? (Hewitt et al., 2019)

1.2.2 Design principles for world–Earth models

To address research questions of the kind suggested by the examples given above, we suggest that the development of WEMs of the type discussed in this paper could be guided by aiming for the following properties.

1. *Explicit representation of social dynamics*. Societal processes should be represented in an explicit, dynamic fashion in order to do justice to the dominant role of human societies in the Anthropocene. (In contrast, social processes occur typically non-dynamically in ESMs as fixed socioeconomic pathways and in IAMs as intertemporal optimization problems.)

   Social processes such as behavioral change as described by the theory of planned behavior (Beckage et al., 2018) or social learning (Donges et al., 2018) may be included in models via comparably simple equation-based descriptions. Yet more detailed WEMs should also allow for representations of the dynamics of the diverse agents and the complex social structure connecting them that constitute human societies, using the tools of agent-based and adaptive network modeling (Farmer and Foley, 2009; Farmer et al., 2015; Müller-Hansen et al., 2017; Lippe et al., 2019; Schill et al., 2019). The social sphere is networked on multiple layers and regarding multiple phenomena (knowledge, trade, institutions, preferences, etc.) and that increasing density of such interacting network structures is one of the defining characteristics of the Anthropocene (Steffen et al., 2007; Gaffney and Steffen, 2017). While there is a rich literature on modeling various aspects of sociocultural dynamics (e.g., Castellano et al., 2009; Snijders et al., 2010; Müller-Hansen et al., 2017; Schlüter et al., 2017), this work so far remains mostly disconnected from Earth system modeling (Calvin and Bond-Lamberty, 2018). Accordingly, more detailed WEMs should be able to describe decision processes of representative samples of individual humans, social groups or classes and collective agents such as firms, households or governments. This includes the representation of diverse objectives, constraints and decision rules, differentiating, for example, by the agent's social class and function and taking the actual and perceived decision options of different agent types into account.

2. *Feedbacks and coevolutionary dynamics*. WEMs should incorporate as dynamic processes the feedbacks of collective social processes on biogeophysical Earth system components and vice versa. The rationale behind this principle is that the strengthening of such feedbacks is one of the key characteristics of the Anthropocene (Beckage et al., 2018; Calvin and Bond-Lamberty, 2018). For example, anthropogenic greenhouse gas emissions drive climate change, which acts back on human societies through increasingly frequent extreme events and may in turn change human behaviors relevant for these emissions. Moreover, the ability to simulate feedbacks is central to a social–ecological and complex adaptive systems approach to Earth system analysis. Capturing these feedbacks enables them to produce paths in coevolution space (Schellnhuber, 1998, 1999) through time-forward integration of all entities and networks allowing for deterministic and stochastic dynamics. Here, time-forward integration refers to simulation of changes in system state over time consecutively in discrete time steps, rather than solving equations that describe the whole time evolution at once as in intertemporal optimization.

3. *Nonlinearity and tipping dynamics*. WEMs should be able to capture the nonlinear dynamics that are a prerequisite for modeling climatic (Lenton et al., 2008; Schellnhuber et al., 2016a; Lenton et al., 2019) and social tipping dynamics (Kopp et al., 2016; Milkoreit et al., 2018; Otto et al., 2020a) and their interactions (Kriegler et al., 2009; Cai et al., 2016) that are not or only partially captured in ESMs and IAMs. This feature is important because the impacts of these critical dynamics are decisive for future trajectories of the Earth system in the Anthropocene, e.g., separating stabilized Earth states that allow for sustainable development from hothouse Earth states of self-amplifying global warming (Heitzig et al., 2016; Steffen et al., 2018).

4. *Cross-scale interactions*. Modeling approaches for investigating social–ecological or coupled human and natural system dynamics have already been developed. However, they usually focus on local or small-scale human–nature interactions (Schlüter et al., 2012). Therefore, such approaches need to be connected across scales and up to the planetary scale and incorporate insights from macro-level and global modeling exercises (Cash et al., 2006; Lippe et al., 2019; Ringsmuth et al., 2019).

5. *Systematic exploration of state and parameter spaces*. WEMs should allow for a comprehensive evaluation of

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

state and parameter spaces to explore the universe of accessible system trajectories and to enable rigorous analyses of uncertainties and model robustness. Hence, they emphasize neither storylines nor optimizations but focus on the exploration of the space of dynamic possibilities to gain systemic understanding. This principle allows for crucial Anthropocene Earth system dynamics to be investigated with state-of-the-art methods from complex systems theory, e.g., for measuring different aspects of the stability and resilience of whole Earth system states and trajectories (Menck et al., 2013; van Kan et al., 2016; Donges and Barfuss, 2017) and for understanding and quantifying planetary boundaries, safe operating spaces, and their manageability and reachability as emergent system properties across scales (Heitzig et al., 2016; Kittel et al., 2017; Anderies et al., 2019).

1.3 World–Earth models compared to existing modeling approaches of global change

It is instructive to compare WEMs more explicitly than above to the two dominant existing classes of global change models – Earth system models and integrated assessment models (van Vuuren et al., 2016) – in terms of the degree to which they represent biophysical, socio-metabolic or economic and sociocultural subsystems and processes in the world–Earth system (Fig. 1). Before discussing how model classes map to these process types, we describe the latter in more detail.

1.3.1 Basic process taxa in world–Earth models

Based on the companion article by Donges et al. (2018) that is also part of the special issue in *Earth System Dynamics* on “Social dynamics and planetary boundaries in Earth system modeling”, we classify processes occurring in the world–Earth system as three major taxa that represent the natural and societal spheres of the Earth system as well as their overlap (Fig. 2). We give only a rough definition and abstain from defining a finer, hierarchical taxonomy, being aware that gaining consensus among different disciplines on such a taxonomy would be unlikely, and we thus leave the assignment of individual processes and attributes to a given taxon to the respective model component developers:

- *Environment (ENV; environmental, biophysical and natural processes).* The “environment” process taxon is meant to contain biophysical or “natural” processes from material subsystems of the Earth system that are not or only insignificantly shaped or designed by human societies (e.g., atmosphere–ocean diffusion, growth of unmanaged vegetation, and maybe the decay of former waste dumps).

- *Metabolism (MET; socio-metabolic and economic processes).* The “metabolism” process taxon is meant to contain socio-metabolic and economic processes from material subsystems that are designed or significantly shaped by human societies (e.g., harvesting, afforestation, greenhouse gas emissions, waste dumping, land-use change, infrastructure building). Social metabolism refers to the material flows in human societies and the way societies organize their exchanges of energy and materials with nature (Fischer-Kowalski, 1997; Martinez-Alier, 2009).

- *Culture (CUL; sociocultural processes).* The “culture” process taxon is meant to contain sociocultural processes from immaterial subsystems (e.g., opinion adoption, social learning, voting, policy making) that are described in models in a way abstracted from their material basis. Culture in its broadest definition refers to everything people do, think and possess as members of society (Bierstedt, 1963, p. 129). Sociocultural processes such as value and norm changes have been suggested to be key for understanding the deeper human dimensions of Earth system dynamics in the Anthropocene (Nyborg et al., 2016; Gerten et al., 2018)

1.3.2 Mapping model classes to Earth system processes

Earth system models focus on the process-detailed description of biogeophysical dynamics (e.g., atmosphere–ocean fluid dynamics or biogeochemistry), while socio-metabolic processes (e.g., economic growth, greenhouse gas emissions and land use) are incorporated via external forcing and sociocultural processes (e.g., public opinion formation, political and institutional dynamics) are only considered implicitly through different scenarios regarding the development of exogenous socio-metabolic drivers (Fig. 1). Integrated assessment models typically contain a more stylized description of biophysical dynamics, are process-detailed in the socio-metabolic or economic domains, and are driven by narratives such as the SSPs (O’Neill et al., 2017) in the sociocultural domain. In turn, WEMs could ultimately integrate all three domains with varying focus depending on the research questions of interest. The focus of current and near-future developments in world–Earth modeling would likely lie on the development of a detailed description of sociocultural processes because they are the ones where the least work has been done so far in formal Earth system modeling.

2 The copan:CORE open world–Earth modeling framework

Here we give a short overview of the world–Earth open modeling framework copan:CORE that was designed following the principles given above (Sect. 1.2) and is more formally described and justified in detail in the Supplement. It enables a flexible model design around standard components

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

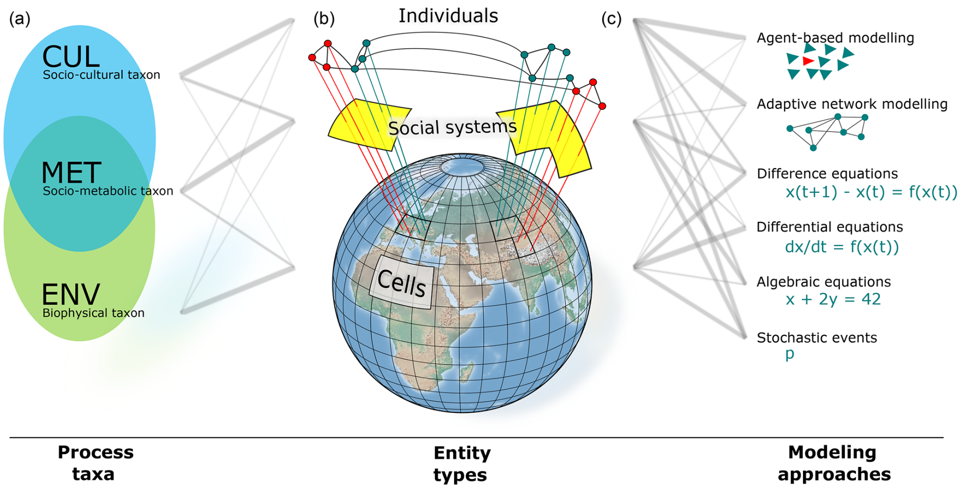


Figure 2. Overview of the copan:CORE open World–Earth modeling framework. The entities in copan:CORE models are classified by entity types (e.g., grid cell, social system, individual; see **b**). Each process belongs to either a certain entity type or a certain process taxon **(a)**. Processes are further distinguished by formal process types (see text for a list), which allow for various different modeling approaches **(c)**. Entity types, process taxa and process types can be freely combined with each other (gray lines). Thick gray lines indicate which combinations are most common. The copan:CORE framework allows us to consistently build world–Earth models across the spectrum from stylized and globally aggregated to more complex and highly resolved variants in terms of spatial and social structure. Hence, entity types, process taxa and types may or may not be present in specific models. For example, a stylized and globally aggregated model would describe the dynamics of the entity types "world" and "social system" and contain neither cells nor individual agents as entities.

and model setups that allows the investigation of a broad set of case studies and research questions using both simple and complex models. Its flexibility and role-based modularization support flexible scripting by end users, interoperability and dynamic coupling with existing models, and a collaborative and structured development in larger teams. copan:CORE is an open, code-based (rather than graphical) simulation modeling framework with a clear focus on Earth system models with endogenous human societies. In other words, it is a tool that provides a standard way of building and running simulation models without giving preference to any particular modeling approach or theory describing human behavior and decision-making and other aspects of social dynamics (Müller-Hansen et al., 2017; Schlüter et al., 2017). Different model components can implement different, sometimes disputed, assumptions about human behavior and social dynamics from theories developed within different fields or schools of thought. This allows for comparison studies in which one component is replaced by a different component modeling the same part of reality in a different way and exploring how the diverging assumptions influence the model outcomes.

All components can be developed and maintained by different model developers and can be flexibly composed into tailor-made models used for particular studies again by different researchers (Fig. 3). The framework facilitates the integration of different types of modeling approaches. It permits, for example, combining micro-economic models (e.g., of a labor market at the level of individuals) with systems of ordinary differential equations (modeling, for example, a carbon cycle). Similarly, systems of implicit and explicit equations (e.g., representing a multi-sector economy) can be combined with Markov jump processes (for example, representing natural hazards). It also provides coupling capabilities to preexisting biophysical Earth system and economic integrated assessment models and thus helps to benefit from the detailed process representations embedded in these models. Many of our design choices are based on experiences very similar to those reported in Robinson et al. (2018), in particular regarding the iterative process of scientific modeling and the need for open code, a common language for a broader community and a high level of consistency without losing flexibility. These features distinguish the copan:CORE modeling framework from existing modeling frameworks and platforms.

A model composed with copan:CORE describes a certain part of the world–Earth system as consisting of a potentially varying set of entities ("things that are", e.g., a spot on the Earth's surface, the European Union, yourself), which are involved in processes ("things that happen", e.g., vegetation growth, economic production, opinion formation) that affect entities' attributes ("how things are", e.g., the spot's harvestable biomass, the EU's gross product, your opinion on fossil fuels, the atmosphere–ocean diffusion coefficient) which represent the variables (including parameters) of a model. An attribute can have a simple or complex data type,

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

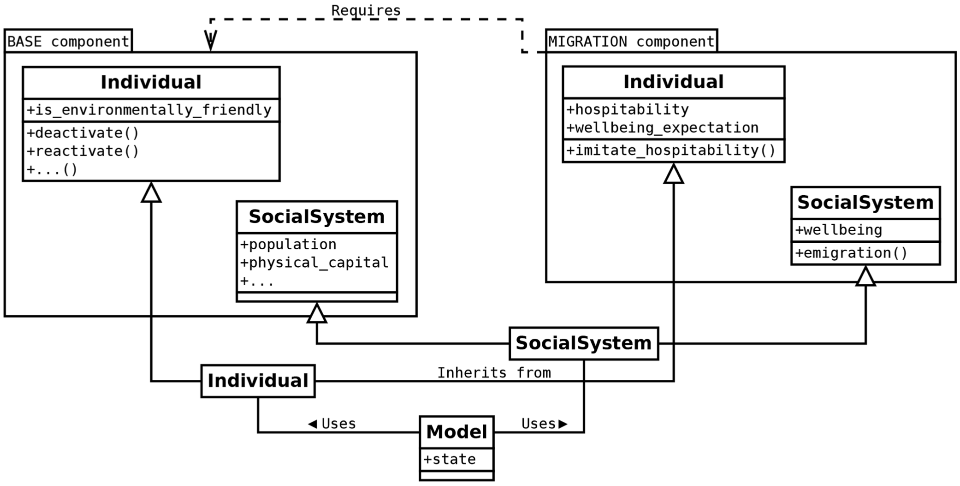


Figure 3. Model composition through multiple inheritance of attributes and processes by process taxa and entity types. This stylized class diagram shows how a model in copan:CORE can be composed from several model components (only two shown here: the mandatory component “base” and the fictitious component “migration”) that contribute component-specific processes and attributes to the model’s process taxa and entity types (only two shown here: “individual” and “SocialSystem”). To achieve this, the classes implementing these entity types on the model level are composed via multiple inheritance (solid arrows) from their component-level counterparts (so-called “mixin” classes).

e.g., representing a binary variable, a whole social network or, to facilitate interoperability and validation, a dimensional quantity with a proper physical unit.

Entities are classified by entity type (cell, social system, individual, etc.), processes by their formal process type (see below), and both are represented by objects in an object-oriented software design, currently using the Python programming language. Each process and each attribute belongs to an entity type or a process taxon (environmental, socio-metabolic, sociocultural). Currently, the following formal process types are supported, enabling typical modeling approaches:

- *ordinary differential equations* representing continuous time dynamics,
- *explicit* or *implicit algebraic equations* representing (quasi-)instantaneous reactions or equilibria,
- *steps* in discrete time representing processes aggregated at the level of some regular time interval or for coupling with external, time-step-based models or model components, and
- *events* happening at irregular or random time points, representing (e.g., agent-based and adaptive network components or externally generated extreme events).

Processes can be implemented either using an imperative programming style via class methods or using symbolic expressions representing mathematical formulae. copan:CORE’s modularization and role concept distinguish between

- *model components* developed by model component developers, implemented as sub-packages of the copan:CORE software package providing interface and implementation mixin classes for entity types and process taxa,
- *models* made from these by model composers, implemented by forming final entity types and process taxa from these mixin classes,
- *studies* by model end users in the form of scripts that import, initialize and run such a model,
- a *master data model* providing metadata for common variables to facilitate interoperability of model components and a common language for modelers, managed by a modeling board.

Entity types and their basic relations shipped with copan:CORE are the following:

- “world”, representing the whole Earth (or some other planet).
- “cell”, representing a regularly or irregularly shaped spatial region used for discretizing the spatial aspect of processes and attributes which are actually continuously distributed in space.

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

- "social system", such as a megacity, country or the EU. It can be interpreted as a human-designed and human-reproduced structure including the flows of energy, material, financial and other resources that are used to satisfy human needs and desires, influenced by the accessibility and use of technology and infrastructure (Fischer-Kowalski, 1997; Otto et al., 2020b), and may include social institutions such as informal systems of norms, values and beliefs and formally codified written laws and regulations, governance, and organizational structures (Williamson, 1998).

- "individual", representing a person, typically used in a network-theoretic, game-theoretic or agent-based component. In contrast to certain economic modeling approaches that use "representative" consumers, an individual in copan:CORE is not meant to represent a whole class of similar individuals (e.g., all the actual individuals of a certain profession) but just one specific individual. Still, the set of all individuals contained in a model will typically be interpreted as being a representative sample of all relevant real-world people. Each individual resides in a cell that belongs to a social system.

Figure 2 illustrates these concepts. Although there is no one-to-one correspondence between process taxa, entity types and modeling approaches, some combinations are expected to occur more often than others, as indicated by the thicker gray connections in Fig. 2. We expect environmental (ENV) processes to deal mostly with cells (for local processes such as terrestrial vegetation dynamics described with spatial resolution) and world(s) (for global processes described without spatial resolution, e.g., the greenhouse effect) and sometimes social systems (for mesoscopic processes described at the level of a social system's territory, e.g., the environmental diffusion and decomposition of industrial wastes). Socio-metabolic (MET) processes will primarily deal with social systems (e.g., for processes described at national or urban level), cells (for local socio-metabolic processes described with additional spatial resolution for easier coupling to natural processes) and world(s) (for global socio-metabolic processes such as international trade) and only rarely with individuals (e.g., for micro-economic model components such as consumption, investment or the job market). Sociocultural (CUL) processes will mostly deal with individuals (for "micro"-level descriptions) and social systems (for "macro"-level descriptions), and rarely world(s) (for international processes such as diplomacy or treaties). Other entity types such as firms, social groups or institutions can be added to the framework if needed.

3 Influence of social dynamics in a minimum-complexity world–Earth model implemented using copan:CORE

In this section, we present an illustrative example of a model realized with our framework. The example model was designed to showcase the concepts and capabilities of copan:CORE in a rather simple WEM, and its components were chosen so that all entity types and process taxa and most features of copan:CORE are covered. Although most model components are somewhat plausible versions of model components that can be found in the various literatures, the example model is intended to be a toy representation of the real world rather than one that could be used directly for studying concrete research questions. Likewise, although we show example trajectories that are based on parameters and initial conditions that roughly reproduce current values of real-world global aggregates in order to make the example as accessible as possible, the time evolutions shown may not be interpreted as any kind of meaningful quantitative prediction or projection.

In spite of this modest goal here, it will become obvious from the presented scenarios that including sociocultural dynamics such as migration, environmental awareness, social learning and policy making in more serious models of the global coevolution of human societies and the environment will likely make a considerable qualitative difference to their results and thus have significant policy implications.

The example model includes the following groups of processes: (1) a version of the simple carbon cycle used in Nitzbon et al. (2017) (based on Anderies et al., 2013) coarsely spatially resolved into four heterogeneous boxes; (2) a version of the simple economy used in Nitzbon et al. (2017) resolved into two world regions. The fossil and biomass energy sectors are complemented by a renewable energy sector with technological progress based on learning by doing (Nagy et al., 2013) and with human capital depreciation; and (3) domestic voting on subsidizing renewables and banning fossil fuels that is driven by individual environmental friendliness. The latter results from becoming aware of environmental problems by observing the local biomass density and diffuses through a social acquaintance network via a standard model of social learning (see, e.g., Holley and Liggett, 1975). These processes cover all possible process taxon interactions as shown in Table 1 and are distributed over six model components in the code as shown in Fig. 4.

We now describe the model components in detail. As many processes add terms to variables' time derivatives, we use the notation $\dot{X} += Y$ to indicate this. The effective time evolution of X is then determined by the sum of the individual processes given below.

Table 1. Possible classification of exemplary model processes by owning process taxon (row) and affected process taxon (column) (following the taxonomy developed in the companion paper Donges et al., 2018): environmental (ENV), social-metabolic (MET) and sociocultural (CUL).

→	CUL	MET	ENV
CUL	social learning, voting	energy policy	environmental protection
MET	well-being	production, capital growth	extraction, harvest, emissions
ENV	well-being, awareness	resource availability	carbon cycle

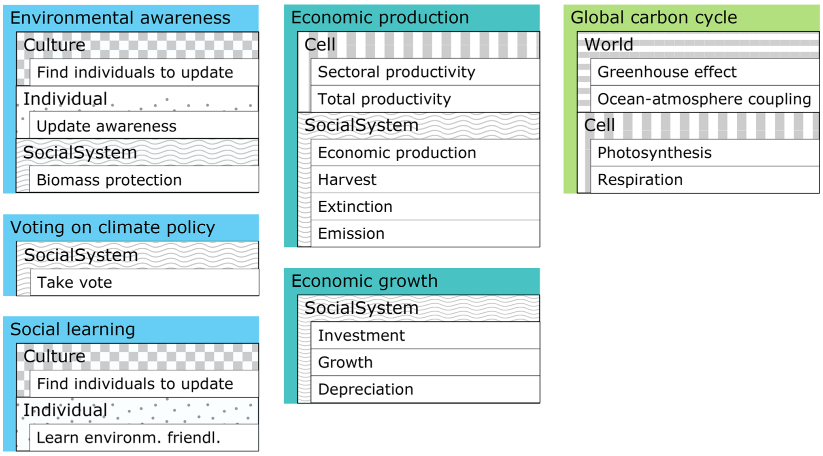


Figure 4. Components, entity types and processes of the example model. Each box represents a model component that contributes several processes (white bars) to different entity types and process taxa (differently hashed rectangles).

3.1 Entity types

The example model contains one world representing the planet, two social systems representing the Global North and South, four cells representing major climate zones: “boreal” and “temperate” belonging to the territory of the Global North and “subtropical” and “tropical” belonging to the Global South, and 100 representative individuals per cell, which form the nodes of a fixed acquaintance network.

3.2 Global carbon cycle

Our carbon cycle follows a simplified version of Anderies et al. (2013) presented in Nitzbon et al. (2017) with coarsely spatially resolved vegetation dynamics. On the world level, an immediate greenhouse effect translates the atmospheric carbon stock A (initially 830 GtC) linearly into a mean surface air temperature $T = T_{\text{ref}} + a(A - A_{\text{ref}})$ (a process of type *explicit equation*) with a sensitivity parameter $a = 1.5\,\text{K}/1000\,\text{GtC}$ and reference values $T_{\text{ref}} = 287\,\text{K}$ and $A_{\text{ref}} = 589\,\text{GtC}$. There is ocean–atmosphere diffusion between A and the upper-ocean carbon stock M (initially 1065 GtC):

$$\dot{A} \mathrel{+}= d(M - mA), \quad \dot{M} \mathrel{+}= d(mA - M) \tag{1}$$

(processes of type “ODE”), with a diffusion rate $d = 0.016\,\text{yr}^{-1}$ and a solubility parameter $m = 1.5$. On the level of a cell c, A and the cell’s terrestrial carbon stock L_c (initially 620 GtC for all four c) are changed by a respiration flow RF_c and a photosynthesis flow PF_c:

$$\dot{A} \mathrel{+}= \text{RF}_c - \text{PF}_c, \quad \dot{L}_c \mathrel{+}= \text{PF}_c - \text{RF}_c. \tag{2}$$

The respiration rate depends linearly on temperature, which is expressed as a dependency on atmospheric carbon density A/Σ, where $\Sigma = 1.5 \times 10^8\,\text{km}^2$ is the total land surface area, so that

$$\text{RF}_c = (a_0 + a_A A/\Sigma) L_c, \tag{3}$$

with a basic rate $a_0 = 0.0298\,\text{yr}^{-1}$ and carbon sensitivity $a_A = 3200\,\text{km}^2\,\text{GtC}^{-1}\,\text{yr}^{-1}$. The photosynthesis rate also depends linearly on temperature (and hence on A) with an additional carbon fertilization factor growing concavely with A/Σ and a space competition factor similar to a logistic equation, giving

$$\text{PF} = (l_0 + l_A A/\Sigma)\sqrt{A/\Sigma}(1 - L_c/k\Sigma_c) L_c, \tag{4}$$

with land area $\Sigma_c = \Sigma/4$, parameters $l_0 = 34\,\text{km}\,\text{GtC}^{-1/2}\,\text{yr}^{-1}$ and $l_A = 1.1 \times 10^6\,\text{km}^3\,\text{GtC}^{-3/2}\,\text{yr}^{-1}$,

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

and per-area terrestrial carbon capacity $k = 25 \times 10^3\,\mathrm{GtC}/1.5 \times 10^8\,\mathrm{km}^2$. Note that the linear temperature dependency and the missing water dependency, in particular, make this model rather stylized; see also Lade et al. (2018).

3.3 Economic production

As in Nitzbon et al. (2017), economic activity consists of producing a final good Y from labor (assumed to be proportional to population P), physical capital K (initially $K_{\mathrm{North}} = 4 \times 10^{13}$, $K_{\mathrm{South}} = 2 \times 10^{13}$, both given in units of USD), and energy input flow E. The latter is the sum of the outputs of three energy sectors, fossil energy flow E_{F}, biomass energy flow E_{B}, and (other) renewable energy flow R. The process is described by a nested Leontieff and Cobb–Douglas production function for Y and Cobb–Douglas production functions for E_{F}, E_{B} and R, all of them here on the level of a cell c:

$$Y_c = y_{\mathrm{E}} \min\left(E_c, b_Y K_{Y,\mathrm{c}}^{\kappa_Y} P_{Y,\mathrm{c}}^{\pi_Y}\right), \quad E_c = E_{\mathrm{F,c}} + E_{\mathrm{B,c}} + R_c, \tag{5}$$

$$E_{\mathrm{F,c}} = b_{\mathrm{F}} K_{\mathrm{F,c}}^{\kappa_{\mathrm{F}}} P_{\mathrm{F,c}}^{\pi_{\mathrm{F}}} G_c^{\gamma}, \tag{6}$$

$$E_{\mathrm{B,c}} = b_{\mathrm{B}} K_{\mathrm{B,c}}^{\kappa_{\mathrm{B}}} P_{\mathrm{B,c}}^{\pi_{\mathrm{B}}} \left(L_c - L_c^{\mathrm{p}}\right)^{\lambda}, \tag{7}$$

$$R_c = b_{\mathrm{R,c}} K_{\mathrm{R,c}}^{\kappa_{\mathrm{R}}} P_{\mathrm{R,c}}^{\pi_{\mathrm{R}}} S_s^{\sigma}. \tag{8}$$

In this, $y_{\mathrm{E}} = \mathrm{USD}\,147\,\mathrm{GJ}^{-1}$ is the energy efficiency, G_c is the cell's fossil reserves (initially 0.4, 0.3, 0.2 and $0.1 \times 1125\,\mathrm{GtC}$ in the boreal, temperate, subtropical and tropical cells), L_c^{p} is the environmentally protected amount of terrestrial carbon (see below), S_s gives the renewable energy production knowledge stock of the corresponding social system s (initially 2×10^{11} GJ), and $\kappa_\bullet = \pi_\bullet = \gamma = \lambda = \sigma = 2/5$ are elasticities leading to slightly increasing returns to scale. The productivity parameters $b_\bullet$ have units that depend on the elasticities and are chosen so that initial global energy flows roughly match the observed values: $b_{\mathrm{F}} = 1.4 \times 10^9\,\mathrm{GJ}^5\,\mathrm{yr}^{-5}\,\mathrm{Gt\,C}^{-2}\,\mathrm{USD}^{-2}$, $b_{\mathrm{B}} = 6.8 \times 10^8\,\mathrm{GJ}^5\,\mathrm{yr}^{-5}\,\mathrm{Gt\,C}^{-2}\,\mathrm{USD}^{-2}$, and $b_{\mathrm{R,c}} = 0.7$, 0.9, 1.1 and 1.3 times the mean value $b_{\mathrm{R}} = 1.75 \times 10^{-11}\,\mathrm{GJ}^3\,\mathrm{yr}^{-5}\,\mathrm{USD}^{-2}$ in boreal, temperate, subtropical and tropical to reflect regional differences in solar insolation. As in Nitzbon et al. (2017), we assume $b_Y \gg b_{\mathrm{B}}$, b_{F}, b_{R} so that its actual value has no influence because then $K_{Y,\mathrm{c}} \ll K_s$ and $P_{Y,\mathrm{c}} \ll Y_s$. Furthermore, $K_{\bullet,\mathrm{c}}$ and $P_{\bullet,\mathrm{c}}$ are the shares of a social system s's capital K_s and labor L_s that are endogenously allocated to the production processes in cell c so that

$$K_s = \sum_{c \in s} \left(K_{Y,\mathrm{c}} + K_{\mathrm{F,c}} + K_{\mathrm{B,c}} + K_{\mathrm{R,c}}\right) \tag{9}$$

and similarly for its population P_s. The latter shares are determined on the social system level in a general equilibrium fashion by equating both wages (marginal productivity of labor) and rents (marginal productivity of capital) in all cells and sectors, assuming costless and immediate labor and capital mobility between all cells and sectors within each social system:

$$\partial y_{\mathrm{E}} E_{\mathrm{F,c}} / \partial P_{\mathrm{F,c}} \equiv \partial y_{\mathrm{E}} E_{\mathrm{B,c}} / \partial P_{\mathrm{B,c}} \equiv \partial y_{\mathrm{E}} R_c / \partial P_{\mathrm{R,c}} \equiv w_s \tag{10}$$

for all $c \in s$, and similarly for $K_{\bullet,\mathrm{c}}$. The production functions and elasticities are chosen so that the corresponding equations can be solved analytically (see Nitzbon et al. (2017) for details), allowing us to first calculate a set of "effective sector or cell productivities" by a process of type *explicit equation* on the cell level, which are used to determine the labor and capital allocation weights $P_{\bullet,\mathrm{c}}/P_s$ and $K_{\bullet,\mathrm{c}}/K_s$, and then calculate output Y_s, carbon emissions, and all cells' fossil and biomass extraction flows in another process of type *explicit equation* on the social system level. Given the latter, a second process of type ODE on the social system level changes the stocks A, G_c and L_c for all cells accordingly.

3.4 Economic growth

Again as in Nitzbon et al. (2017), but here on the social system level, a fixed share i (here 0.244) of economic production Y_s is invested into physical capital K_s:

$$\dot{K}_s \mathrel{+}= \; i Y_s. \tag{11}$$

Capital also depreciates at a rate that depends linearly on surface air temperature to represent damage from climate change:

$$\dot{K}_s \mathrel{+}= \; -(k_0 + k_T (T - T_K)) K_s \tag{12}$$

with $k_0 = 0.1\,\mathrm{yr}^{-1}$, $k_T = 0.05\,\mathrm{yr}^{-1}\,\mathrm{K}^{-1}$, and $T_K = 287$ K. In addition, renewable energy production knowledge S_s grows proportional to its utilization via learning by doing:

$$\dot{S}_s \mathrel{+}= \; R_s. \tag{13}$$

Finally, we interpret S_s as a form of human capital that also depreciates at a constant rate (due to forgetting or becoming useless because of changing technology, etc.):

$$\dot{S}_s \mathrel{+}= \; -\beta S_s, \tag{14}$$

with $\beta = 0.02\,\mathrm{yr}^{-1}$. Note that unlike in Nitzbon et al. (2017), we consider populations to be constant at $P_{\mathrm{North}} = 1.5 \times 10^9$ and $P_{\mathrm{South}} = 4.5 \times 10^9$ to avoid the complexities of a well-being-driven population dynamics component (which could, however, be implemented in the same way as in Nitzbon et al. (2017) on the social system level).

3.5 Environmental awareness

On the level of the culture process taxon, an "awareness updating" process of type "event" occurs at random time points with a constant rate (i.e., as a Poisson process, here with a rate of $4\,\mathrm{yr}^{-1}$), representing times at which many people become aware of the state of the environment, e.g., because of notable environmental events. At each such a time point, each

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

individual independently updates their environmental friendliness (a Boolean variable) with a certain probability. When individuals update, they switch from “false” to “true” with a probability ψ^+ depending on the terrestrial carbon density in their cell c, $\mathrm{TCD}_c = L_c/\Sigma_c$, given by

$$\psi^+ = \exp\left(-\mathrm{TCD}_c/\mathrm{TCD}^\perp\right), \quad (15)$$

and switches from true to false with a probability

$$\psi^- = 1 - \exp\left(-\mathrm{TCD}_c/\mathrm{TCD}^\top\right), \quad (16)$$

where $\mathrm{TCD}^\perp = 1 \times 10^{-5}$ and $\mathrm{TCD}^\top = 4 \times 10^{-5}$ are sensitivity parameters with $\mathrm{TCD}^\perp < \mathrm{TCD}^\top$ to generate hysteresis behavior. As a consequence, a fraction L_c^P of the terrestrial carbon L_c is protected from harvesting for economic production. This fraction is proportional to the cell’s social system’s population share represented by those individuals which are environmentally friendly. The initial share of environmentally friendly individuals will be varied in the bifurcation analysis below.

3.6 Social learning

Similarly, on the culture level, “social learning” events occur at random time points with a constant rate (here $4\,\mathrm{yr}^{-1}$), representing times at which the state of the environment becomes a main topic in the public debate. At each such time point, each individual i independently compares their environment with that of a randomly chosen acquaintance j with a certain fixed probability (here 1/10). j then convinces i to copy j’s environmental friendliness with a probability ψ that depends via a sigmoidal function on the difference in logs between both home cells’ terrestrial carbon densities:

$$\psi = 1/2 + \arctan\left(\pi\phi'\left(\log \mathrm{TCD}_j - \log \mathrm{TCD}_i - \log\rho'\right)\right)/\pi, \quad (17)$$

where $\phi' = 1$ and $\rho' = 1$ are slope and offset parameters. The underlying social network is a block model network in which each individual is on average linked to 10 randomly chosen others: 5 in the same cell, 3.5 in the other cell of the same social system and 1.5 in the other social system.

3.7 Voting on climate policy

Each (of the two) social systems performs general elections at regular time intervals (hence implemented as a process of type “step”, here every 4 years) which may lead to the introduction or termination of climate policies. If at the time t of the election, more than a certain threshold (here 1/2) of the population is environmentally friendly, both a subsidy for renewables (here USD $50\,\mathrm{GJ}^{-1}$) is introduced and use of fossils is banned. This leads to a shift in the energy price equilibrium that determines the energy sector’s allocation of labor and capital, which then reads

$$\begin{aligned}&\text{marginal production cost of biomass energy}\\&= \text{marginal production cost of renewable energy}\\&- \text{renewable subsidy.}\end{aligned}$$

Conversely, if these policies are already in place but the environmentally friendly population share is below some other thresholds (here also 1/2), these policies are terminated.

Note that we have chosen to model awareness formation and social learning in an agent-based fashion here mainly to illustrate that such an approach can easily be combined with other approaches in copan:CORE, not because we want to claim that an agent-based approach is the most suitable here. Indeed, one may well want to replace these two agent-based model components by equation-based versions which approximate their behavior in terms of macroscopic quantities (e.g., as in Wiedermann et al., 2015), and because of the modular design of copan:CORE, this can easily be done and the two model versions could be compared (nevertheless, this is beyond the scope of this paper).

3.8 Results

In order to show in particular what effect the inclusion of sociocultural processes into WEMs can have on their results, we compare two representative 100-year runs of the example model described above: one without the sociocultural processes environmental awareness, social learning, and voting (left panels of Fig. 5) and another with these processes included (right panels of Fig. 5). Both runs start in model year 0 from the same initial conditions and use the same parameters, which were chosen to roughly reflect real-world global aggregates of the year 2000 (see above). For the simulation without social processes (left panels of Fig. 5) both social systems (“Global North” as solid and “Global South” as dashed lines) initially rely on fossil energy in order to meet their energy needs, thus causing a rise in atmospheric and ocean carbon and a decline in fossil carbon stocks. Similarly both social systems initially rely heavily on energy from biomass, with the consequence of a reduction in terrestrial carbon. Due to the technology becoming competitive, the Global South changes its energy production to renewable energy comparatively early in the simulation, resulting in a fast fading out of biomass and fossils as an energy source. Due to its larger fossil reserves and lower solar insolation, the Global North takes 2 decades longer to make this switch. However, this delay in the Global North causes high atmospheric carbon, hence a high global mean temperature, which due to our oversimplified vegetation model makes the terrestrial carbon stock decline further even after biomass has been phased out as an energy source as well, recovering only much later (not shown). In both social systems, economic growth declines until the switch, then boosts and later declines again since neither population nor total factor productivity grow in our model. Once the Global South switches to renewables, it hence overtakes the Global North, and this reversed inequality is then sustained as our model includes no trade, knowledge spillovers, migration or other direct interaction which would lead to economic convergence. Certainly, such results are not in themselves realistic (as this model does not intend to be) or transferable to real-world application. Future WEMs, therefore, should include such processes beyond pure economic ones in order to properly capture real-world–Earth dynamics; see the Supplement for some corresponding extensions of this model.

If social processes are considered, we obtain qualitatively similar but quantitatively different trajectories, e.g., in the right panels of Fig. 5, where we initially assume 40 % of all individuals are environmentally friendly. As before, both social systems initially rely on energy produced from fossils and biomass, but as biomass reduces terrestrial carbon density, environmental awareness makes some people environmentally friendly and this spreads via social learning. Once half of the population is environmentally friendly, the next elections in that social system bring a fossil ban and subsidies for renewables. This causes a slightly earlier switch to renewables than before, especially in the Global North (dashed lines in Fig. 5). This ultimately results in lower atmospheric and ocean carbon stocks, lower peak temperatures, less cumulative use of fossil fuels and a much faster recovery of terrestrial carbon.

copan:CORE further allows for a systematic investigation of the influence of individual parameters on the outcome of the simulation (e.g., along the lines of a bifurcation analysis). As an illustration of such an analysis we now vary the learning rate from $1/50\,\mathrm{yr}^{-1}$ (less than once in a generation) to $12\,\mathrm{yr}^{-1}$ (once every month) and compute the carbon stocks as well as the GDP per capita and the global mean temperature in model year 120 for an ensemble of 50 simulations per learning rate (Fig. 6) and the same initial conditions for all runs (we thus do not test for a possible multistability of the system).

For learning rates lower than $1\,\mathrm{yr}^{-1}$ (slow learning) the carbon stocks as well as the global mean temperature align well for the two simulation setups, i.e., the one with (scatter points) and without social processes (dashed lines). In contrast, for learning rates larger than $1\,\mathrm{yr}^{-1}$ (faster learning) the individuals become more capable of assessing the consequences of their behavior (in our case extensive biomass use) before the system has reached a state with low terrestrial and high atmospheric and ocean carbon stocks. As such, increasing the learning rate also causes an increase in the terrestrial carbon stock combined with a decrease in the atmospheric and ocean carbon stocks (in model year 120). This behavior is also reflected in the global mean temperature which decreases as the learning rate increases. Hence, with respect to the environment, social learning only has a positive effect if it happens at a sufficiently high rate (around once to more than once a year). It remains to note that learning rates have in the past already been shown to have a profound impact on the state and dynamics of a coupled socio-ecological system, a feature that is recovered in our simple WEM as well (Wiedermann et al., 2015; Auer et al., 2015; Barfuss et al., 2017).

The metabolic variable GDP per capita interestingly already increases much earlier (i.e., for much lower learning rates than $1\,\mathrm{yr}^{-1}$) as compared to the changes in the environmental variables. This implies that for our specific WEM, social processes generally seem to foster the economy regardless of their actual rate. Furthermore we observe that the Global South shows an approximately 3 times higher GDP per capita than the Global North, which is caused by the earlier switch to renewable energies in that social system (see third row of Fig. 5). As already stated above, note again, that these results are not intended as a realistic projection of future trajectories of the Earth system, but are discussed here to showcase the capabilities of the copan:CORE framework.

Using the pycopancore reference implementation, running the above two simulations (Fig. 5) took 140 s (without sociocultural processes) and 290 s (including sociocultural processes) on an Intel Xeon E5-2690 CPU at 2.60 GHz. Since further performance improvements are desirable to support Monte Carlo simulations, we aim at a community-supported development of an alternative, more production-oriented implementation in the C++ language.

4 Conclusions

In this paper, we presented a simulation modeling framework that aims at facilitating the implementation and analysis of world–Earth (or planetary social–ecological) models. It follows a modular design such that various model components can be combined in a plug-and-play fashion to easily explore the influence of specific processes or the effect of competing theories of social dynamics from different schools of thought (Schlüter et al., 2017) on the coevolutionary trajectories of the system. The model components describe fine-grained yet meaningfully defined subsystems of the social and environmental domains of the world–Earth system and thus enable the combination and comparison of different modeling approaches from the natural and social sciences. In the modeling framework, different entities such as geographic cells, individual humans and social systems are represented and their attributes are shaped by environmental, socio-metabolic and sociocultural processes. The mathematical types of processes that can be implemented in the modeling framework range from ordinary differential and algebraic equations to deterministic and stochastic events. Due to its flexibility, the model framework can be used to analyze interactions at and between various scales – from local to regional and global.

The current version of the copan:CORE open modeling framework includes a number of tentative model components implementing, e.g., basic economic, climatic, biological, demographic and social network dynamics. However,

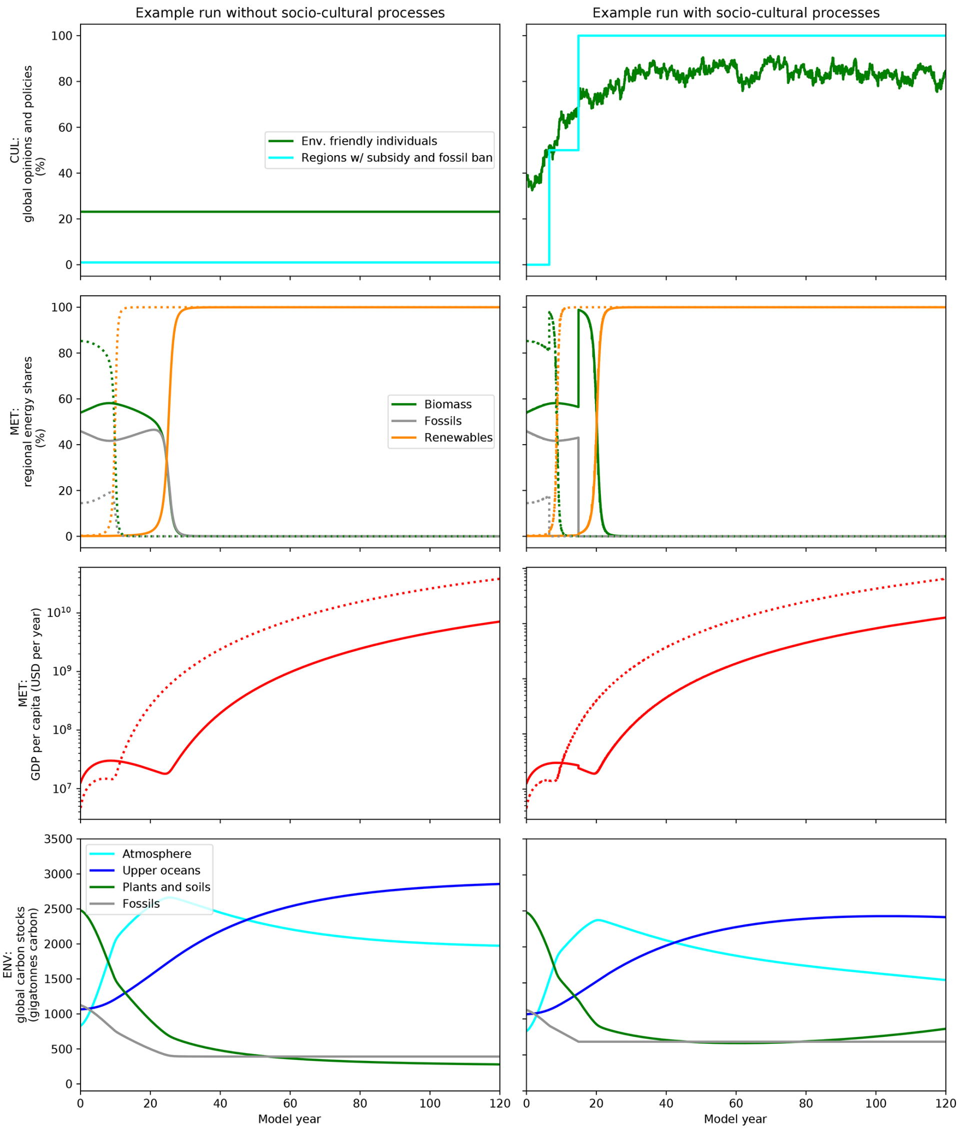


Figure 5. Two runs from a world–Earth model example: one without (left panels) and one with (right panels) the sociocultural processes of environmental awareness, social learning and voting included, showing different transient (and asymptotic) behavior. The top row shows variables related to the cultural process taxon, the second and third row those related to the metabolic process taxon and the bottom row those related to the environmental process taxon. Green, orange, cyan, blue and gray lines correspond to variables related to terrestrial carbon, renewables, atmospheric carbon, ocean carbon and fossils, respectively. In second and third row, dashed lines indicate variables associated with the "Global South", solid lines to the "Global North".

to use the modeling framework for rigorous scientific analyses, these components have to be refined, their details have to be spelled out and new components have to be developed that capture processes with crucial influence on world–Earth coevolutionary dynamics. For this purpose, various modeling approaches from the social sciences are available to be applied to develop comprehensive representations of such socio-metabolic and sociocultural processes (Müller-Hansen et al., 2017; Schill et al., 2019, and references therein). For example, hierarchical adaptive network approaches could be used to model the development of social groups, institutions and organizations spanning local to global scales or the interaction of economic sectors via resource, energy and infor-

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

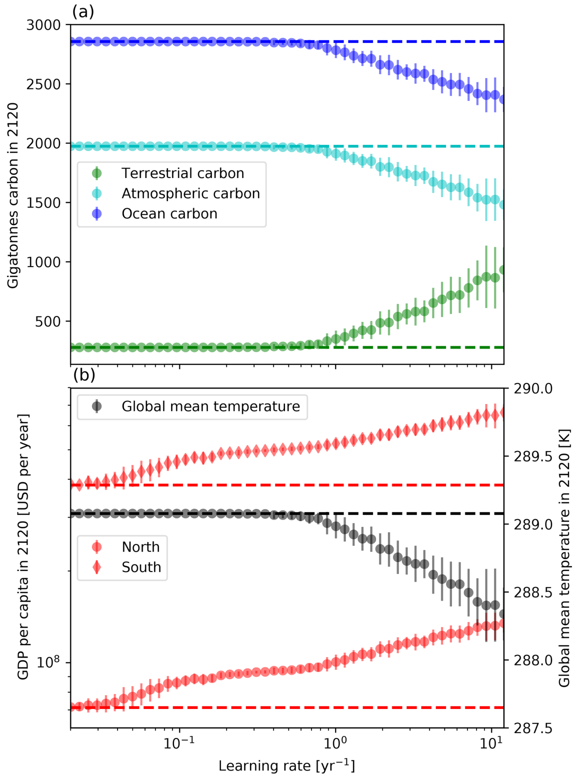


Figure 6. Dependency of some selected variables after 120 model years on the learning rate of environmental awareness. Scatter points denote (the average over 50) simulations with social processes, and error bars denote 1 standard deviation for each choice of learning rate. Dashed lines indicate the corresponding values for a simulation without social processes. Panel **(a)** shows the three environmental (non-fossil) carbon stocks; panel **(b)** shows the GDP per capita in the two social systems as well as the global mean temperature.

mation flows (Gross and Blasius, 2008; Donges et al., 2017a; Geier et al., 2019).

Making such an endeavor prosper requires the collection and synthesis of knowledge from various disciplines. The modular approach of the copan:CORE open modeling framework supports well-founded development of single model components, helps to integrate various processes and allows analyzing their interplay. To facilitate this, we envision an emergent community of modelers who contribute mature model components, composed models and variable definitions that add to a growing master component and model repository, and a master data model that are hosted within the open-source software repository (see below under "Code availability"), curated by a repository management board and cross-linked with platforms such as the CoMSES network (https://www.comses.net, last access: 1 April 2020). Complete models should also be contributed. This way, copan:CORE could support the emergence of community standards for modeling coupled human–natural systems that have recently been demanded by many researchers (Barton and The Open Modeling Foundation, 2019). We therefore call upon the interdisciplinary social–ecological modeling community and beyond to participate in further model and application development to facilitate "whole" Earth system analysis of the Anthropocene.

Code availability. A Python 3.7.x implementation of the copan:CORE open World–Earth modeling framework, detailed documentation, a tutorial and the world–Earth model example are available at https://doi.org/10.5281/zenodo.3772751 (Heitzig et al., 2020).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/esd-11-395-2020-supplement.

Author contributions. JFD and JH designed and coordinated the study. MW and JH performed model simulations and analyzed results. All other authors contributed to the writing of the paper and the discussion of results.

Competing interests. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue "Social dynamics and planetary boundaries in Earth system modelling". It is not associated with a conference.

Acknowledgements. This work has been carried out within the framework of PIK's flagship project on Coevolutionary Pathways in the Earth system (copan, https://www.pik-potsdam.de/research/projects/activities/copan/copan-introduction, last access: 1 April 2020). We are grateful for financial support by the Stordalen Foundation via the Planetary Boundary Research Network (PB.net), the Earth League's EarthDoc program, the Leibniz Association (project DominoES), the European Research Council (ERC advanced grant project ERA), and the German Federal Ministry of Education and Research (BMBF, project CoNDyNet). We acknowledge additional support by the Heinrich Böll Foundation (WB), the Foundation of German Business (JJK), the Episcopal Scholarship Foundation Cusanuswerk (JK) and DFG/FAPESP (IRTG 1740/TRP 2015/50122-0, FMH). The European Regional Development Fund, BMBF and the Land Brandenburg supported this project by providing resources on the high-performance computer system at the Potsdam Institute for Climate Impact Research. We thank the participants of the three LOOPS workshops (https://www.pik-potsdam.de/research/projects/activities/copan/loops/loops-workshop-series, last access: 1 April 2020) in Kloster Chorin (2014), Southampton (2015) and Potsdam (2017) for discussions that provided highly valuable in-

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

sights for conceptualizing world–Earth modeling and the development of the copan:CORE open simulation modeling framework.

Financial support. The article processing charges for this open-access publication were covered by the Potsdam Institute for Climate Impact Research (PIK).

Review statement. This paper was edited by James Dyke and reviewed by Brian Dermody, Carsten Lemmen, and Axel Kleidon.

References

Anderies, J. M., Carpenter, S. R., Steffen, W., and Rockström, J.: The topology of non-linear global carbon dynamics: from tipping points to planetary boundaries, Environ. Res. Lett., 8, 44048, https://doi.org/10.1088/1748-9326/8/4/044048, 2013.

Anderies, J. M., Mathias, J.-D., and Janssen, M. A.: Knowledge infrastructure and safe operating spaces in social–ecological systems, P. Natl. Acad. Sci. USA, 116, 5277–5284, 2019.

Arneth, A., Brown, C., and Rounsevell, M.: Global models of human decision-making for land-based mitigation and adaptation assessment, Nat. Clim. Change, 4, 550–557, 2014.

Auer, S., Heitzig, J., Kornek, U., Schöll, E., and Kurths, J.: The dynamics of coalition formation on complex networks, Scient. Rep., 5, 13386, https://doi.org/10.1038/srep13386, 2015.

Barfuss, W., Donges, J. F., Wiedermann, M., and Lucht, W.: Sustainable use of renewable resources in a stylized social–ecological network model under heterogeneous resource distribution, Earth Syst. Dynam., 8, 255–264, https://doi.org/10.5194/esd-8-255-2017, 2017.

Barfuss, W., Donges, J. F., Lade, S. J., and Kurths, J.: When optimization for governing human-environment tipping elements is neither sustainable nor safe, Nat. Commun., 9, 2354, https://doi.org/10.1038/s41467-018-04738-z, 2018.

Barton, M. and The Open Modeling Foundation: An Open Letter in Support of Community Standards for Modeling Coupled Human-Natural Systems, available at: https://openmodelingfoundation.org/open-letter (last access: 11 January 2020), 2019.

Beckage, B., Gross, L. J., Lacasse, K., Carr, E., Metcalf, S. S., Winter, J. M., Howe, P. D., Fefferman, N., Franck, T., Zia, A., Kinzig, A., and Hoffman, F. M.: Linking models of human behaviour and climate alters projected climate change, Nat. Clim. Change, 8, 79–84, 2018.

Berkes, F., Folke, C., and Colding, J.: Linking social and ecological systems: management practices and social mechanisms for building resilience, Cambridge University Press, Cambridge, 2000.

Betts, R. K.: Conflict after the Cold War: arguments on causes of war and peace, Taylor & Francis, New York, 2017.

Bierstedt, R.: The Social Order, McGraw-Hill, New York, 1963.

Cai, Y., Lenton, T. M., and Lontzek, T. S.: Risk of multiple interacting tipping points should encourage rapid CO_2 emission reduction, Nat. Clim. Change, 6, 520–525, 2016.

Calvin, K. and Bond-Lamberty, B.: Integrated human-earth system modeling – state of the science and future directions, Environ. Res. Lett., 13, 063006, https://doi.org/10.1088/1748-9326/aac642, 2018.

Cash, D., Adger, W. N., Berkes, F., Garden, P., Lebel, L., Olsson, P., Pritchard, L., and Young, O.: Scale and cross-scale dynamics: governance and information in a multilevel world, Ecol. Soc., 11, https://doi.org/10.5751/ES-01759-110208, 2006.

Castellano, C., Fortunato, S., and Loreto, V.: Statistical physics of social dynamics, Rev. Mod. Phys., 81, 591–646, https://doi.org/10.1103/RevModPhys.81.591, 2009.

Crutzen, P. J.: Geology of mankind, Nature, 415, 23–23, 2002.

Cumming, G. S. and Peterson, G. D.: Unifying research on social–ecological resilience and collapse, Trends Ecol. Evol., 32, 695–713, 2017.

Dermody, B. J., Sivapalan, M., Stehfest, E., van Vuuren, D. P., Wassen, M. J., Bierkens, M. F. P., and Dekker, S. C.: A framework for modelling the complexities of food and water security under globalisation, Earth Syst. Dynam., 9, 103–118, https://doi.org/10.5194/esd-9-103-2018, 2018.

Di Baldassarre, G., Martinez, F., Kalantari, Z., and Viglione, A.: Drought and flood in the Anthropocene: feedback mechanisms in reservoir operation, Earth Syst. Dynam., 8, 225–233, https://doi.org/10.5194/esd-8-225-2017, 2017.

Donges, J. F. and Barfuss, W.: From Math to Metaphors and Back Again: Social-Ecological Resilience from a Multi-Agent-Environment Perspective, GAIA – Ecol. Perspect. Sci. Soc., 26, 182–190, 2017.

Donges, J. F., Lucht, W., Müller-Hansen, F., and Steffen, W.: The technosphere in Earth System analysis: A coevolutionary perspective, Anthropocene Rev., 4, 23–33, 2017a.

Donges, J. F., Winkelmann, R., Lucht, W., Cornell, S. E., Dyke, J. G., Rockström, J., Heitzig, J., and Schellnhuber, H. J.: Closing the loop: Reconnecting human dynamics to Earth System science, Anthropocene Rev., 4, 151–157, 2017b.

Donges, J. F., Lucht, W., Heitzig, J., Barfuss, W., Cornell, S. E., Lade, S. J., and Schlüter, M.: Taxonomies for structuring models for World–Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops, Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2018-27, in review, 2018.

Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Farahani, E., Kadner, S., Seyboth, K., Adler, A., Baum, I., Brunner, S., Eickemeier, P., Kriemann, B., Savolainen, J., Schlömer, S., von Stechow, C., Zwickel, T., and Minx, J. (Eds.): Climate Change 2014: Mitigation of Climate Change, in: Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK and New York, NY, USA, 2014.

Edenhofer, O., Flachsland, C., Jakob, M., and Lessmann, K.: The atmosphere as a global commons, in: The Oxford Handbook of the Macroeconomics of Global Warming, edited by: Bernard, L. and Semmler, W., Oxford University Press, New York, NY, USA, 2015.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Farmer, J. D. and Foley, D.: The economy needs agent-based modelling, Nature, 460, 685–686, 2009.

Farmer, J. D., Hepburn, C., Mealy, P., and Teytelboym, A.: A third wave in the economics of climate change, Environ. Resour. Econ., 62, 329–357, 2015.

Farmer, J. D., Hepburn, C., Ives, M., Hale, T., Wetzer, T., Mealy, P., Rafaty, R., Srivastav, S., and Way, R.: Sensitive intervention points in the post-carbon transition, Science, 364, 132–134, 2019.

Fischer-Kowalski, M.: On the Childhood and Adolescence of a Rising Conceptual Star, in: The International Handbook of Environmental Sociology, Edward Elgar Publishing, Cheltenham, UK, 1997.

Foley, A. M., Holden, P. B., Edwards, N. R., Mercure, J.-F., Salas, P., Pollitt, H., and Chewpreecha, U.: Climate model emulation in an integrated assessment framework: a case study for mitigation policies in the electricity sector, Earth Syst. Dynam., 7, 119–132, https://doi.org/10.5194/esd-7-119-2016, 2016.

Folke, C.: Resilience: The emergence of a perspective for social–ecological systems analyses, Global Environ. Change, 16, 253–267, 2006.

Future Earth: Future Earth Strategic Research Agenda 2014, International Council for Science (ICSU), Paris, 2014.

Gaffney, O. and Steffen, W.: The Anthropocene equation, Anthropocene Rev., 4, 53–61, 2017.

Garrett, T. J.: Long-run evolution of the global economy: 1. Physical basis, Earth's Future, 2, 127–151, 2014.

Geier, F., Barfuss, W., Wiedermann, M., Kurths, J., and Donges, J. F.: The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources, Eur. Phys. J. Spec. Top., 228, 2357–2369, 2019.

Gerten, D., Schönfeld, M., and Schauberger, B.: On deeper human dimensions in Earth system analysis and modelling, Earth Syst. Dynam., 9, 849–863, https://doi.org/10.5194/esd-9-849-2018, 2018.

Gross, T. and Blasius, B.: Adaptive coevolutionary networks: a review, J. Roy. Soc. Interface, 5, 259–271, 2008.

Haff, P.: Humans and technology in the Anthropocene: Six rules, Anthropocene Rev., 1, 126–136, 2014.

Haff, P. K.: Technology and human purpose: the problem of solids transport on the Earth's surface, Earth Syst. Dynam., 3, 149–156, https://doi.org/10.5194/esd-3-149-2012, 2012.

Heck, V., Donges, J. F., and Lucht, W.: Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dynam., 7, 783–796, https://doi.org/10.5194/esd-7-783-2016, 2016.

Heitzig, J. and Kornek, U.: Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility, Nat. Clim. Change, 8, 204–209, https://doi.org/10.1038/s41558-018-0079-z, 2018.

Heitzig, J., Kittel, T., Donges, J. F., and Molkenthin, N.: Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dynam., 7, 21–50, https://doi.org/10.5194/esd-7-21-2016, 2016.

Heitzig, J., Donges, J. F., Barfuss, W., Breitbach, P., Kassel, J., Kittel, T., Kolster, T., Kolb, J., Müller-Hansen, F., Wiedermann, M., and Zimmerer, K.: pycopancore – Reference implementation of the copan:CORE World–Earth modeling framework, https://doi.org/10.5281/zenodo.3772751, 2020.

Helbing, D., Bishop, S., Conte, R., Lukowicz, P., and McCarthy, J. B.: FuturICT: Participatory computing to understand and manage our complex world in a more sustainable and resilient way, Eur. Phys. J. Spec. Top., 214, 11–39, 2012.

Hewitt, R., Bradley, N., Baggio Compagnucci, A., Barlagne, C., Ceglarz, A., Cremades, R., McKeen, M., Otto, I., and Slee, B.: Social Innovation in Community Energy in Europe: A Review of the Evidence, Front. Energ. Res., 7, 31, https://doi.org/10.3389/fenrg.2019.00031, 2019.

Holley, R. A. and Liggett, T. M.: Ergodic theorems for weakly interacting infinite systems and the voter model, Ann. Probabil., 3, 643–663, 1975.

Jager, W., Janssen, M., De Vries, H., De Greef, J., and Vlek, C.: Behaviour in commons dilemmas: Homo economicus and Homo psychologicus in an ecological-economic model, Ecol. Econ., 35, 357–379, 2000.

Janssen, M.: Complexity and ecosystem management: the theory and practice of multi-agent systems, Edward Elgar Publishing, Cheltenham, UK/Northampton, MA, USA, 2002.

Janssen, M. and De Vries, B.: The battle of perspectives: a multi-agent model with adaptive responses to climate change, Ecol. Econ., 26, 43–65, 1998.

Kellie-Smith, O. and Cox, P. M.: Emergent dynamics of the climate–economy system in the Anthropocene, Philos. T. Roy. Soc. A, 369, 868–886, 2011.

Keys, P. W. and Wang-Erlandsson, L.: On the social dynamics of moisture recycling, Earth Syst. Dynam., 9, 829–847, https://doi.org/10.5194/esd-9-829-2018, 2018.

Kittel, T., Koch, R., Heitzig, J., Deffuant, G., Mathias, J.-D., and Kurths, J.: Operationalization of Topology of Sustainable Management to Estimate Qualitatively Different Regions in State Space, arXiv preprint, arXiv:1706.04542, 2017.

Kopp, R. E., Shwom, R. L., Wagner, G., and Yuan, J.: Tipping elements and climate–economic shocks: Pathways toward integrated assessment, Earth's Future, 4, 346–372, 2016.

Kriegler, E., Hall, J. W., Held, H., Dawson, R., and Schellnhuber, H. J.: Imprecise probability assessment of tipping points in the climate system, P. Natl. Acad. Sci. USA, 106, 5041–5046, 2009.

Lade, S. J., Donges, J. F., Fetzer, I., Anderies, J. M., Beer, C., Cornell, S. E., Gasser, T., Norberg, J., Richardson, K., Rockström, J., and Steffen, W.: Analytically tractable climate–carbon cycle feedbacks under 21st century anthropogenic forcing, Earth Syst. Dynam., 9, 507–523, https://doi.org/10.5194/esd-9-507-2018, 2018.

Lenton, T. M. and Latour, B.: Gaia 2.0, Science, 361, 1066–1068, 2018.

Lenton, T. M. and Watson, A. J.: Revolutions that made the Earth, Oxford University Press, Oxford, 2011.

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J.: Tipping elements in the Earth's climate system, P. Natl. Acad. Sci. USA, 105, 1786–1793, 2008.

Lenton, T. M., Pichler, P.-P., and Weisz, H.: Revolutions in energy input and material cycling in Earth history and human history, Earth Syst. Dynam., 7, 353–370, https://doi.org/10.5194/esd-7-353-2016, 2016.

Lenton, T. M., Rockström, J., Gaffney, O., Rahmstorf, S., Richardson, K., Steffen, W., and Schellnhuber, H.: Climate tipping points-too risky to bet against, Nature, 575, 592–595, 2019.

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

Lippe, M., Bithell, M., Gotts, N., Natalini, D., Barbrook-Johnson, P., Giupponi, C., Hallier, M., Hofstede, G. J., Le Page, C., Matthews, R. B., Schlüter, M., Smith, P., Teglio, A., and Thellmann, K.: Using agent-based modelling to simulate social-ecological systems across scales, GeoInformatica, 23, 269–298, 2019.

Liu, J., Dietz, T., Carpenter, S. R., Alberti, M., Folke, C., Moran, E., Pell, A. N., Deadman, P., Kratz, T., Lubchenco, J., Ostrom, E., Ouyang, Z., Provencher, W., Redman, C. L., Schneider, S. H., and Taylor, W. W.: Complexity of coupled human and natural systems, Science, 317, 1513–1516, 2007.

Lucht, W. and Pachauri, R.: Earth system analysis for sustainability, in: chap. The mental component of the Earth system, MIT Press, Cambridge, MA, 341–365, 2004.

Martinez-Alier, J.: Social metabolism, ecological distribution conflicts, and languages of valuation, Capital. Nat. Social., 20, 58–87, 2009.

Meadows, D., Meadows, D., Randers, J., Behrens III, W., Hakimzadeh, F., Harbordt, S., Machen, J. A., Meadows, D., Milling, P., Murthy, N. S., Naill, R. F., Randers, J., Shantzis, S., Seeger, J. A., Williams, M., and Zahn, E. K. O.: The limits to growth, in: A report for the Club of Rome's project on the predicament of mankind, Universe Books, New York, 1972.

Menck, P. J., Heitzig, J., Marwan, N., and Kurths, J.: How basin stability complements the linear-stability paradigm, Nat. Phys., 9, 89–92, 2013.

Milkoreit, M., Hodbod, J., Baggio, J., Benessaiah, K., Calderón-Contreras, R., Donges, J. F., Mathias, J.-D., Rocha, J. C., Schoon, M., and Werners, S. E.: Defining tipping points for social-ecological systems scholarship – an interdisciplinary literature review, Environ. Res. Lett., 13, 033005, https://doi.org/10.1088/1748-9326/aaaa75, 2018.

Motesharrei, S., Rivas, J., and Kalnay, E.: Human and Nature Dynamics (HANDY): Modeling inequality and use of resources in the collapse or sustainability of societies, Ecol. Econ., 101, 90–102, 2014.

Müller-Hansen, F., Schlüter, M., Mäs, M., Donges, J. F., Kolb, J. J., Thonicke, K., and Heitzig, J.: Towards representing human behavior and decision making in Earth system models – an overview of techniques and approaches, Earth Syst. Dynam., 8, 977–1007, https://doi.org/10.5194/esd-8-977-2017, 2017.

Nagy, B., Farmer, J. D., Bui, Q. M., and Trancik, J. E.: Statistical Basis for Predicting Technological Progress, PLoS ONE, 8, 1–7, https://doi.org/10.1371/journal.pone.0052669, 2013.

Nitzbon, J., Heitzig, J., and Parlitz, U.: Sustainability, collapse and oscillations in a simple World-Earth model, Environ. Res. Lett., 12, 074020, https://doi.org/10.1088/1748-9326/aa7581, 2017.

Nyborg, K., Anderies, J. M., Dannenberg, A., Lindahl, T., Schill, C., Schlüter, M., Adger, W. N., Arrow, K. J., Barrett, S., Carpenter, S., Chapin III, F. S., Crépin, A. S., Daily, G., Ehrlich, P., Folke, C., Jager, W., Kautsky, N., Levin, S. A., Madsen, O. J., Polasky, S., Scheffer, M., Walker, B., Weber, E. U., Wilen, J., Xepapadeas, A., and de Zeeuw, A.: Social norms as solutions, Science, 354, 42–43, 2016.

O'Neill, B. C., Kriegler, E., Ebi, K. L., Kemp-Benedict, E., Riahi, K., Rothman, D. S., van Ruijven, B. J., van Vuuren, D. P., Birkmann, J., Kok, K., Levy, M., and Solecki, W.: The roads ahead: Narratives for shared socioeconomic pathways describing world futures in the 21st century, Global Environ. Change, 42, 169–180, 2017.

O'Neill, D. W., Fanning, A. L., Lamb, W. F., and Steinberger, J. K.: A good life for all within planetary boundaries, Nat. Sustainabil., 1, 88–95, 2018.

Ostrom, E.: Governing the commons: The evolution of institutions for collective action, Cambridge University Press, Cambridge, 1990.

Otto, I. M., Biewald, A., Coumou, D., Feulner, G., Köhler, C., Nocke, T., Blok, A., Gröber, A., Selchow, S., Tyfield, D., Volkmer, I., Schellnhuber, H. J., and Beck, U.: Socio-economic data for global environmental change research, Nat. Clim. Change, 5, 503–506, 2015.

Otto, I. M., Reckien, D., Reyer, C. P., Marcus, R., Le Masson, V., Jones, L., Norton, A., and Serdeczny, O.: Social vulnerability to climate change: a review of concepts and evidence, Reg. Environ. Change, 17, 1651–1662, 2017.

Otto, I. M., Kim, K. M., Dubrovsky, N., and Lucht, W.: Shift the focus from the super-poor to the super-rich, Nat. Clim. Change, 9, 82–94, 2019.

Otto, I. M., Donges, J. F., Cremades, R., Bhowmik, A., Hewitt, R. J., Lucht, W., Rockström, J., Allerberger, F., McCaffrey, M., Doe, S. P. S., Lenferna, A., Moran, N., van Vuuren, D. P., and Schellnhuber, H. J.: Social tipping dynamics for stabilizing Earth's climate by 2050, P. Natl. Acad. Sci. USA, 117, 2354–2365, 2020a.

Otto, I. M., Wiedermann, M., Cremades, R., Donges, J. F., Auer, C., and Lucht, W.: Human agency in the Anthropocene, Ecol. Econ., 167, 106463, https://doi.org/10.1016/j.ecolecon.2019.106463, 2020b.

Raworth, K.: A safe and just space for humanity: can we live within the doughnut, Oxfam Policy Pract.: Clim. Change Resilience, 8, 1–26, 2012.

Ringsmuth, A. K., Lade, S. J., and Schlüter, M.: Cross-scale cooperation enables sustainable use of a common-pool resource, P. Roy. Soc. B, 286, 20191943, https://doi.org/10.1098/rspb.2019.1943, 2019.

Robinson, D. T., Di Vittorio, A., Alexander, P., Arneth, A., Michael Barton, C., Brown, D. G., Kettner, A., Lemmen, C., O'Neill, B. C., Janssen, M., Pugh, T. A., Rabin, S. S., Rounsevell, M., Syvitski, J. P., Ullah, I., and Verburg, P. H.: Modelling feedbacks between human and natural processes in the land system, Earth Syst. Dynam., 9, 895–914, https://doi.org/10.5194/esd-9-895-2018, 2018.

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E. F., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R. W., Fabry, V. J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., and Foley, J. A.: A safe operating space for humanity, Nature, 461, 472–475, 2009.

Rockström, J., Brasseur, G., Hoskins, B., Lucht, W., Schellnhuber, J., Kabat, P., Nakicenovic, N., Gong, P., Schlosser, P., Máñez Costa, M., Humble, A., Eyre, N., Gleick, P., James, R., Lucena, A., Masera, O., Moench, M., Schaeffer, R., Seitzinger, S., van der Leeuw, S., Ward, B., Stern, N., Hurrell, J., Srivastava, L., Morgan, J., Nobre, C., Sokona, Y., Cremades, R., Roth, E., Liverman, D., and Arnott, J.: Climate change: The necessary, the possible and the desirable Earth League climate statement on the implications for climate policy from the 5th IPCC Assessment, Earth's Future, 2, 606–611, 2014.

Rounsevell, M. D. A., Arneth, A., Alexander, P., Brown, D. G., de Noblet-Ducoudré, N., Ellis, E., Finnigan, J., Galvin, K., Grigg, N., Harman, I., Lennox, J., Magliocca, N., Parker, D., O'Neill, B. C., Verburg, P. H., and Young, O.: Towards decision-based global land use models for improved understanding of the Earth system, Earth Syst. Dynam., 5, 117–137, https://doi.org/10.5194/esd-5-117-2014, 2014.

Schellnhuber, H. J.: Discourse: Earth System analysis – The scope of the challenge, in: Earth System Analysis, Springer, Berlin, Heidelberg, 3–195, 1998.

Schellnhuber, H. J.: Earth system analysis and the second Copernican revolution, Nature, 402, C19–C23, 1999.

Schellnhuber, H. J., Rahmstorf, S., and Winkelmann, R.: Why the right climate target was agreed in Paris, Nat. Clim. Change, 6, 649–653, 2016a.

Schellnhuber, H. J., Serdeczny, O., Adams, S., Köhler, C., Otto, I., and Schleussner, C.: The Challenge of a 4 Degrees Celcius World by 2100, in: Handbook on Sustainability Transition and Sustainable Peace, Hexagon Series on Human and Environmental Security and Peace, vol 10, edited by: Brauch, H. G., Oswald Spring, U., Grin, J., and Scheffran, J., Springer, Cham, 2016b.

Schill, C., Anderies, J. M., Lindahl, T., Folke, C., Polasky, S., Cárdenas, J. C., Crépin, A.-S., Janssen, M. A., Norberg, J., and Schlüter, M.: A more dynamic understanding of human behaviour for the Anthropocene, Nat. Sustainabil., 2, 1075–1082, 2019.

Schlüter, M., McAllister, R. R. L., Arlinghaus, R., Bunnefeld, N., Eisenack, K., Hölker, F., and Milner-Gulland, E. J.: New horizons for managing the environment: A review of coupled social-ecological systems modeling, Nat. Resour. Model., 25, 219–272, 2012.

Schlüter, M., Baeza, A., Dressler, G., Frank, K., Groeneveld, J., Jager, W., Janssen, M. A., McAllister, R. R., Müller, B., Orach, K., Schwarz, N., and Wijermans, N.: A framework for mapping and comparing behavioural theories in models of social-ecological systems, Ecol. Econ., 131, 21–35, 2017.

Snijders, T. A., Van de Bunt, G. G., and Steglich, C. E.: Introduction to stochastic actor-based models for network dynamics, Social Netw., 32, 44–60, 2010.

Steffen, W., Crutzen, P. J., and McNeill, J. R.: The Anthropocene: are humans now overwhelming the great forces of nature, Ambio, 36, 614–621, 2007.

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., de Vries, W., de Wit, C. A., Folke, C., Gerten, D., Heinke, J., Mace, G. M., Persson, L. M., Ramanathan, V., Reyers, B., and Sörlin, S.: Planetary boundaries: Guiding human development on a changing planet, Science, 347, 736–747, doi10.1126/science.1259855, 2015.

Steffen, W., Rockström, J., Richardson, K., Lenton, T., Folke, C., Liverman, D., Summerhayes, C., Barnosky, A., Cornell, S., Crucifix, M., Donges, J., Fetzer, I., Lade, S., Scheffer, M., Winkelmann, R., and Schellnhuber, H.: Trajectories of the Earth system in the Anthropocene, P. Natl. Acad. Sci. USA, 115, 8252–8259, 2018.

Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. (Eds.): Climate Change 2013: The Physical Science Basis, in: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK and New York, NY, USA, 2013.

Tàbara, J. D., Frantzeskaki, N., Hölscher, K., Pedde, S., Kok, K., Lamperti, F., Christensen, J. H., Jäger, J., and Berry, P.: Positive tipping points in a rapidly warming world, Curr. Opin. Environ. Sustainabil., 31, 120–129, 2018.

van Kan, A., Jegminat, J., Donges, J. F., and Kurths, J.: Constrained basin stability for studying transient phenomena in dynamical systems, Phys. Rev. E, 93, 042205, https://doi.org/10.1103/PhysRevE.93.042205, 2016.

van Vuuren, D. P., Bayer, L. B., Chuwah, C., Ganzeveld, L., Hazeleger, W., van den Hurk, B., Van Noije, T., O'Neill, B., and Strengers, B. J.: A comprehensive view on climate change: coupling of earth system and integrated assessment models, Environ. Res. Lett., 7, 024012, https://doi.org/10.1088/1748-9326/7/2/024012, 2012.

van Vuuren, D. P., Lucas, P. L., Häyhä, T., Cornell, S. E., and Stafford-Smith, M.: Horses for courses: analytical tools to explore planetary boundaries, Earth Syst. Dynam., 7, 267–279, https://doi.org/10.5194/esd-7-267-2016, 2016.

Verburg, P. H., Dearing, J. A., Dyke, J. G., van der Leeuw, S., Seitzinger, S., Steffen, W., and Syvitski, J.: Methods and approaches to modelling the Anthropocene, Global Environ. Change, 39, 328–340, 2016.

Wiedermann, M., Donges, J. F., Heitzig, J., Lucht, W., and Kurths, J.: Macroscopic description of complex adaptive networks coevolving with dynamic node states, Phys. Rev. E, 91, 052801, https://doi.org/10.1103/PhysRevE.91.052801, 2015.

Wiedermann, M., Winkelmann, R., Donges, J. F., Eder, C., Heitzig, J., Katsanidou, A., and Smith, E. K.: Domino Effects in the Earth System – The potential role of wanted tipping points, preprint arXiv:1911.10063 [physics.soc-ph], 2019.

Willett, W., Rockström, J., Loken, B., Springmann, M., Lang, T., Vermeulen, S., Garnett, T., Tilman, D., DeClerck, F., Wood, A., Jonell, M., Clark, M., Gordon, L. J., Fanzo, J., Hawkes, C., Zurayk, R., Rivera, J. A., De Vries, W., Sibanda, L. M., Afshin, A., Chaudhary, A., Herrero, M., Agustina, R., Branca, F., Lartey, A., Fan, S., Crona, B., Fox, E., Bignet, V., Troell, M., Lindahl, T., Singh, S., Cornell, S. E., Reddy, K. S., Narain, S., Nishtar, S., and Murray, C. J. L.: Food in the Anthropocene: the EAT – Lancet Commission on healthy diets from sustainable food systems, Lancet, 393, 447–492, 2019.

Williamson, O. E.: Transaction cost economics: how it works; where it is headed, Economist, 146, 23–58, 1998.

Reproduced from: J. F. Donges et al., Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework, Earth Syst. Dyn., vol. 11, no. 2, pp. 395-413, 2020, doi: 10.5194/esd-11-395-2020. Published under Creative Commons Attribution License 4.0 (CC BY).

3
Theoretical and methodological work

This third section is dedicated to theoretical publications on World-Earth models and to present new analytical methods for the analysis of complex social-ecological systems, which is structured in four subsections.

In "Stability and resilience of complex social-ecological systems" (Sect. 3.1) we present a selection of our conceptual and theoretical work as a basis for further investigations of complex social-ecological systems. In particular, we describe investigations of system resilience through nonlinear stability analyses.

The subsequent section on "Sustainable management of complex social-ecological systems" (Sect. 3.2) deals with qualitative differences between regions in systems' state spaces and their connectivity with respect to sustainable management. Furthermore, we introduce methodologies for analysing sustainable management options in social-ecological systems.

In the third section on "Dynamics of adaptive social-ecological networks" (Sect. 3.3), we focus on the characteristic dynamics of social-ecological systems described as complex networks. Two exemplary models serve as showcases.

We close with "Model simplification and approximation methods" (Sect. 3.4) by showing approaches for describing emerging macroscopic phenomena in agent-based social-ecological dynamics on networks that help getting a deeper understanding of the behaviour of these complex systems.

3.1 *Stability and resilience of complex social-ecological systems*

This section presents selected analyses of resilience and stability of complex social-ecological systems.

The concept of "basin stability", introduced earlier [Menck et al., 2013], builds on and formalises influential conceptualisations of social-ecological resilience [Holling, 1973] and has widely been used to study the sensitivity of complex systems. In "Constrained basin stability for studying transient phenomena in dynamical systems" [Van Kan et al., 2016], we introduced a generalization of this that takes also transient behaviour into account.

Adding to this, in "Survivability of deterministic dynamical systems" [Hellmann et al., 2016] we joined our colleagues from the COEN project in defining survivability in a complex system as the likelihood that the transient behaviour of a deterministic system does not leave a region of desirable states.

Trajectories might converge to a stable attractor but still take infinitely long to get there. The paper "Timing of transients: quantifying reaching times and transient behavior in complex systems" [Kittel et al., 2017a] focuses on the question of how long such a transient phase takes.

We close with a proposal for a classification of modern notions of social-ecological resilience from a multi-agent-environment perspective in "From math to metaphors and back again: social-ecological resilience from a multi-agent-environment perspective" [Donges and Barfuss, 2017]. Building on this, we demonstrated why a further development of the mathematics of resilience and advancing models suitable for the study of social-ecological-technological system resilience is necessary [Tamberg et al., 2020].

PHYSICAL REVIEW E **93**, 042205 (2016)

Constrained basin stability for studying transient phenomena in dynamical systems

Adrian van Kan,[1,2,*] Jannes Jegminat,[1,†] Jonathan F. Donges,[3,4] and Jürgen Kurths[3,5,6,7]

[1] *Department of Physics and Astronomy, University of Heidelberg, Im Neuenheimer Feld 226, D-69120 Heidelberg, Germany*
[2] *Department of Physics, Imperial College London, Prince Consort Rd, London SW7 2BB, United Kingdom*
[3] *Potsdam Institute for Climate Impact Research, P.O. Box 601203, D-14412 Potsdam, Germany*
[4] *Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden*
[5] *Department of Physics, Humboldt University Berlin, Newtonstr. 15, D-12489 Berlin, Germany*
[6] *Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3FX, United Kingdom*
[7] *Department of Control Theory, Nizhny Novgorod State University, Gagarin Avenue 23, 606950 Nizhny Novgorod, Russia*

(Received 13 August 2015; revised manuscript received 9 January 2016; published 13 April 2016)

Transient dynamics are of large interest in many areas of science. Here, a generalization of basin stability (BS) is presented: constrained basin stability (CBS) that is sensitive to various different types of transients arising from finite size perturbations. CBS is applied to the paradigmatic Lorenz system for uncovering nonlinear precursory phenomena of a boundary crisis bifurcation. Further, CBS is used in a model of the Earth's carbon cycle as a return time-dependent stability measure of the system's global attractor. Both case studies illustrate how CBS's sensitivity to transients complements BS in its function as an early warning signal and as a stability measure. CBS is broadly applicable in systems where transients matter, from physics and engineering to sustainability science. Thus CBS complements stability analysis with BS as well as classical linear stability analysis and will be a useful tool for many applications.

DOI: 10.1103/PhysRevE.93.042205

I. INTRODUCTION

Many fields of science analyze dissipative dynamical systems in terms of their attractors. Thus it is an important challenge to quantify the stability of attractors with respect to a given perturbation. The most popular method is linear stability analysis, which considers infinitesimal perturbations. Menck *et al.* [1] suggest to complement this linear measure with basin stability (BS), which accounts for finite and even large perturbations. The application of BS to power grids has yielded novel mitigation strategies against superoutages [2,3]. BS is computed by estimating the volume of an attractor's basin. Therefore, it is not sensitive to different forms of transient dynamics. However, transient phenomena in complex systems are of large interest in many areas of science, such as climatic and, more generally, global change in Earth system science [4], epileptic seizures in neuroscience [5], ecosystem transitions in ecology [6], as well as in the previously mentioned study of super outages in power grids [2,3]. For example, in the case of climate change [4] and the great acceleration [7] as transient phenomena in the global social-environmental system [8,9], major efforts are invested into studying the maximum global mean temperature and its timing along the trajectory due to anthropogenic greenhouse gas emissions. Moreover, the model- and data-driven analysis of transient global change trajectories underlies many recently proposed frameworks for sustainable development such as tolerable environment and development windows [10], planetary boundaries [11,12], and the safe and just operating space for humanity [13].

Making BS sensitive to transients, we generalize it to a family of stability measures termed constrained basin stabilities (CBSs). As opposed to BS, CBS is not computed from the entire basin of an attractor but only from a subset of the basin. The subset is defined by a generic constraint imposed on the transients. Thus CBS is sensitive to transients while maintaining the intuitiveness and simplicity of BS. To illustrate how CBS complements BS, we choose two specific constraints on transients, one based on the confinement of transient trajectories to certain regions in phase space and one based on transient duration, and apply them to the Lorenz system and a global carbon cycle model, respectively. In the former example, CBS anticipates a boundary crisis bifurcation. In the latter, we show that CBS represents a more intuitive measure for stability than BS because CBS reflects not only that perturbation-induced transients return to the attractor but also that they do so within a desirable time interval.

This paper is structured as follows. In Sec. II we introduce CBS and discuss some of its properties used in the further analysis. In Sec. III we present two examples of CBS analysis in dynamical systems: the paradigmatic Lorenz [14] model and a global carbon cycle model proposed by Anderies *et al.* [15]. Then, in Sec. IV we discuss the relevance of CBS and how it differs from established stability concepts. The paper concludes with closing remarks.

II. METHODS

Let the (not necessarily analytic) vector valued function f represent an autonomous potentially multistable dynamical system $\dot{\mathbf{x}} = f(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^n$, and let $\phi^t(\mathbf{x}_0)$ be the system state at time t on a trajectory starting at $\mathbf{x}_0$ at $t = 0$. The basin stability $S_B(A)$ of an attractor A of this system quantifies the probability that after a finite size perturbation trajectories return to A [1]. Perturbations within the attractor's basin $\mathcal{B}(A)$ return but the remaining ones fall into a different attractor. The probability distribution of perturbing a trajectory on the attractor to the state $\mathbf{x}$ is given by $\rho(\mathbf{x})$. Thereby, BS is formally

[*] van_kan@stud.uni-heidelberg.de
[†] jegminat@iup.uni-heidelberg.de

Reprinted paper with permission from A. Van Kan, J. Jegminat, J. F. Donges, and J. Kurths, Constrained basin stability for studying transient phenomena in dynamical systems, Phys. Rev. E, vol. 93, no. 4, pp. 1-7, 2016, doi: 10.1103/PhysRevE.93.042205. Copyright 2016 by the American Physical Society.

defined as

$$S_B(A) = \int_\Gamma dx^n \rho(\mathbf{x}) \mathbf{1}_{\mathcal{B}(A)}(\mathbf{x}), \tag{1}$$

where Γ denotes the state or phase space of the dynamical system and the indicator function is

$$\mathbf{1}_{\mathcal{B}(A)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in \mathcal{B}(A), \\ 0 & \text{else.} \end{cases} \tag{2}$$

BS can be computed quickly once an attractor's basin is given: it equals the mass of ρ that is supported by the basin. However, in practice, the basin of attraction is usually not known and needs to be determined first. For this purpose, initial conditions are sampled according to the perturbation density ρ and then integrated until they reach an attractor. Thus, in computing BS only the long-term limit of the trajectories is used to determine if an initial condition lies in the basin of attraction. Therefore, by construction, BS does not depend on transient motion.

To generalize BS, here we propose instead to use the properties of transients to define a class of stability measures that we term *constrained basin stabilities* (CBSs). Calculating CBS requires a computational effort similar to that needed for BS, but CBS reveals additional information about the system that is encoded in the transient trajectories. We define a transient as the set of points belonging to the part of the trajectory between the initial condition $\mathbf{x}(0) = \mathbf{x}_0$ and reaching the attractor A,

$$T(\mathbf{x}_0) = \{\phi^t(\mathbf{x}_0) \in \Gamma \backslash A \mid t \geqslant 0\}. \tag{3}$$

The fact that we define the attractor not to be part of the transient makes a difference for example in the case of trajectories induced by nonsmooth flows where an attractor may be reached within finite time. The idea of CBS is that a region in phase space is identified by some constraint on the transients starting from a subset of phase space

$$C = \{\mathbf{x} \in \Gamma \backslash A | \text{ the transient from } \mathbf{x} \text{ satisfies a constraint}\}. \tag{4}$$

In other words, the transients starting from this *conditioned set* C satisfy the given constraint. For instance, we can choose C to be the set of states $\mathbf{x}$ the transients starting from which exhibit monotonicity in the x_1 component. This is equivalent to demanding that the projection of a transient's velocity onto the basis vector $\mathbf{e}_1$ in x_1 direction is nonvanishing. Thus the conditioned set is $C_{mon} = \{\mathbf{x} \in \Gamma \backslash A | f(\phi^t(\mathbf{x})) \cdot \mathbf{e}_1 \neq 0 \,\forall\, t > 0\}$. If x_1 represents a population, the set C_{mon} is the set of initial conditions which do not lead to a population overshoot [16].

Incorporating an arbitrary constraint (not necessarily monotonicity) as an additional factor in Eq. (1), we formally define CBS as

$$S_B^C(A) = \int_\Gamma dx^n \rho(\mathbf{x}) \mathbf{1}_C(\mathbf{x}) \mathbf{1}_{\mathcal{B}(A)}(\mathbf{x}). \tag{5}$$

The product of the two indicator functions checks whether a perturbation is inside of the attractor's basin and at the same time inside the region transients originating from which satisfy the prescribed constraint. Figure 1 illustrates the regions in a schematic two-dimensional phase space that are relevant for computing BS and CBS for a fixed point. Three useful properties of CBS follow directly from its definition. First, since $\mathbf{1}_C(\mathbf{x}) \leqslant 1$,

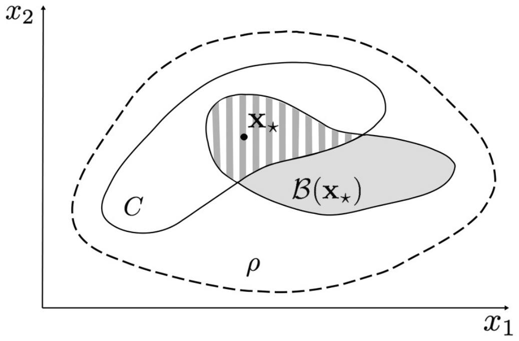


FIG. 1. Subspaces of a schematic two-dimensional phase space containing a fixed point $\mathbf{x}_\star$: perturbations are sampled from the domain of the perturbation probability density ρ (area within dashed line). Their transients return to the fixed point when sampled from the basin of attraction $\mathcal{B}(\mathbf{x}_\star)$ (gray area). The set C (white within solid line) is the set of initial conditions which lead to transients that satisfy a given constraint, e.g., on the x_1 component. While BS is computed as the fraction of perturbations within the basin of attraction, CBS is the fraction of perturbations within the intersection $C \cap \mathcal{B}(\mathbf{x}_\star)$ (striped area). Thus CBS reflects the stability with respect to perturbations whose transients fulfill a given constraint.

$$S_B(A) \geqslant S_B^C(A). \tag{6}$$

Secondly, let $\{C_i\}_{i \in I}$ be a partition of Γ, then

$$\sum_{i \in I} S_B^{C_i}(A) = S_B(A). \tag{7}$$

Thirdly, if $C_1 \subset C_2$ then

$$S_B^{C_1}(A) \leqslant S_B^{C_2}(A). \tag{8}$$

The novelty of CBSs is that they integrate information about the transients into the asymptotic framework of BS. This information is encoded in a set C of states the transients originating from which satisfy a given requirement, such as monotonicity in the previous example. We suggest classifying these requirements as static, dynamic, and integrated, depending on how much information is necessary to find out which desirable region of phase space corresponds to them as follows. (i) Static requirements *a priori* define a phase space region $\Gamma' \subset \Gamma$ that must not be entered by the transient. No further knowledge of the system or its dynamics is required. Examples are planetary boundaries in Earth system dynamics [11,12], minimum or maximum operating temperatures of a device or the evaluation of external functions (not f) of the system, e.g., the performance of a second system that depends on the system state $\mathbf{x}$. (ii) Dynamic requirements depend on velocity, thus more knowledge about the system is required: the dynamics, i.e., f, must be known. Using this knowledge, a region similar to Γ' is defined. Consider, for example, a roller coaster that must not exceed a certain velocity or acceleration or the requirement of monotonicity in economic output to exclude the burst of market bubbles. (iii) Integrated conditions depend not only on the current state of the transient but also on its past,

Reprinted paper with permission from A. Van Kan, J. Jegminat, J. F. Donges, and J. Kurths, Constrained basin stability for studying transient phenomena in dynamical systems, Phys. Rev. E, vol. 93, no. 4, pp. 1-7, 2016, doi: 10.1103/PhysRevE.93.042205. Copyright 2016 by the American Physical Society.

i.e., they operate on an infinite dimensional space and memory effects are possible. Despite this complexity, testing integrated conditions is often easy in practice as can be seen from the following examples: imposing a limited number of opinion changes of a political party, thresholding the time needed to reach an attractor, integrated damage in a climate model or imposing a minimum average power output of a wind farm per time interval. Note that each of the constraints implies a binary decision: CBS identifies a qualitative property in a transient.

In order to implement CBS [Eq. (5)] numerically, we need to discretize it. For simplicity of presentation, we choose the attractor to be a fixed point, $A = \{\mathbf{x}_\star\}$, and consider a uniform distribution ρ of N initial conditions drawn from some subset of phase space approximated by a set of sampling points $\mathbf{x}_i$, $i \in \{1,\dots,N\}$ drawn at random from the phase space volume in question. This results in

$$S_B^C(\mathbf{x}_\star,\varepsilon) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_C(\mathbf{x}_i)\, \Theta(\varepsilon - d_{\min}), \tag{9}$$

where $d_{\min}$ is the minimal state-space distance (within the finite simulation time) between the fixed point $\mathbf{x}_\star$ and the transient $T(\mathbf{x}_i)$ and $\Theta(x)$ is the Heaviside function.

If a trajectory reaches a distance smaller than the threshold ε from the attractor within finite simulation time, we regard it to have reached the attractor. Furthermore, the uniformity of ρ implies $\rho(\mathbf{x}_i) = N^{-1}$. Operationally, for any attractor A (not necessarily a fixed point), we proceed as follows: (1) sample an initial condition $\mathbf{x}_i$ according to ρ; (2) integrate $\mathbf{x}_i$ in time until it has reached an attractor; (3) if the reached attractor is A, count $\mathbf{x}_i$ towards $S_B(A)$; (4) check if the transient originating from $\mathbf{x}_i$ satisfies the constraint, if so count $\mathbf{x}_i$ towards $S_B^C(A)$; (5) increase $i \to i+1$; repeat until $i = N$.

The computational procedure outlined above by which we determine BS and CBS allows us to estimate the uncertainty of our estimates of BS and CBS. Since we consider a uniform perturbation ρ, we are effectively drawing initial conditions at random from the subset R of phase space where ρ is nonzero. The fraction p of the volume the basin $\mathcal{B}(A)$ occupied by R is the true BS, i.e., the probability that we draw an initial condition from $\mathcal{B}(A)$ at random. This implies that, effectively, our estimate of BS after drawing N initial conditions comes from a binomial distribution with expectation value $p = S_B$, which leads to the standard deviation

$$\sigma_{S_B(A)} = \frac{1}{N}\sqrt{S_B(1-S_B)N} = \frac{1}{\sqrt{N}}\sqrt{S_B(1-S_B)}. \tag{10}$$

Equation (10) also holds when $S_B(A)$ is replaced by $S_B^C(A)$, which follows from an argument analogous to the one above. It is important at this point to note that nonuniform distributions ρ are also admissible and make sense in certain applications when some perturbations need to be weighted more than others. However, an error estimate as in Eq. (10) is less straightforward to obtain for nonuniform ρ.

III. APPLICATION

To illustrate the versatility of CBS, we give examples of specific constraints in the paradigmatic Lorenz system [14] and in a global carbon cycle model by Anderies *et al.* [15]. In the Lorenz63 (L63) model we show how CBS can reveal precursory phenomena before the onset of a boundary crisis bifurcation. In the Anderies model, we argue that CBS reflects our intuition of stability of a desired state against perturbations better than standard BS. We illustrate in both examples how CBS can generate important new insights into the dynamics of complex systems, while being simple enough to be amenable to a quick interpretation.

A. Anticipating a boundary crisis bifurcation

The L63 system [14]

$$\dot{x} = \sigma(y - x), \tag{11}$$

$$\dot{y} = rx - y - xz, \tag{12}$$

$$\dot{z} = xy - bz \tag{13}$$

is a conceptual model of Rayleigh-Bénard convection. It is famous for exhibiting chaotic dynamics along with a rich dynamical behavior. Setting $\sigma = 10$ and $b = 8/3$, we begin by summarizing the bifurcation structure as the parameter $r \in [9,26]$ increases. At first, two stable fixed points exist at $\mathbf{x}_\star^{(\pm)} = (\pm\sqrt{b(r-1)}, \pm\sqrt{b(r-1)}, r-1)$, corresponding to left and right turning convection rolls, respectively. At $r_1 = 13.926$ a chaotic saddle appears. At $r_2 = 24.06$ this chaotic saddle undergoes a boundary crisis and becomes attractive. The fixed points lose their stability at $r_3 = 24.74$. Our goal is to anticipate this boundary crisis [17]. To this end, we choose a specific condition sensitive to long (chaotic) transients, since these are precursors of the crisis. With the following static constraint, we discriminate between transients that stay close to one of the fixed points $\mathbf{x}_+$ or $\mathbf{x}_-$ (i.e., one sense of convective overturning) and chaotic transients that flip between them:

$$C^{\pm} = \{\mathbf{x} \in \Gamma \mid \phi^t(\mathbf{x}) \cdot \mathbf{n} \neq 0\ \forall t > 0\}, \tag{14}$$

where the difference vector $\mathbf{n} = |\mathbf{x}_\star^+ - \mathbf{x}_\star^-|^{-1}(\mathbf{x}_\star^+ - \mathbf{x}_\star^-)$ is the normal of a plane H containing the origin that separates the phase space into two symmetric halves. Figure 2 shows two-dimensional cross sections of the three-dimensional basins of attraction $\mathcal{B}(\mathbf{x}_*^{\pm})$ and their intersections $C^{\pm} \cap \mathcal{B}(\mathbf{x}_*^{\pm})$ with the sets $C^{\pm}$ defined above for two different values of the parameter r.

BS and CBS are computed for 15 625 initial conditions sampled from the uniform perturbation distribution $\rho(\mathbf{x}) = \frac{1}{40}\Theta(20-|x|)\frac{1}{40}\Theta(20-|y|)\frac{1}{30}\Theta(15-|z|)$, which describes a box that roughly covers the attractor. To compute BS, we evolve each initial condition in time until it reaches either one of the fixed points or the chaotic attractor. The termination condition in the former case is that the trajectory enters an ε-ball around $\mathbf{x}_\star^{\pm}$; here $\varepsilon = 10^{-4}$. In the latter case, we iterate until the trajectory has crossed the plane H a large (but computationally feasible) number m of times; here $m = 400$. Figure 3 shows the resulting BS and CBS. The BS and CBS curves are identical for both fixed points due to their symmetry. Thus only the values for the positive fixed point are shown. For any conditioned set C and its complement $\overline{C} = \Gamma \backslash C$, Eq. (7) reduces to $S_B(A) = S_B^C(A) + S_B^{\overline{C}}(A)$. This implies that the difference between the two stability measures reflects the fraction of the basin from where (long) chaotic transients

Reprinted paper with permission from A. Van Kan, J. Jegminat, J. F. Donges, and J. Kurths, Constrained basin stability for studying transient phenomena in dynamical systems, Phys. Rev. E, vol. 93, no. 4, pp. 1-7, 2016, doi: 10.1103/PhysRevE.93.042205. Copyright 2016 by the American Physical Society.

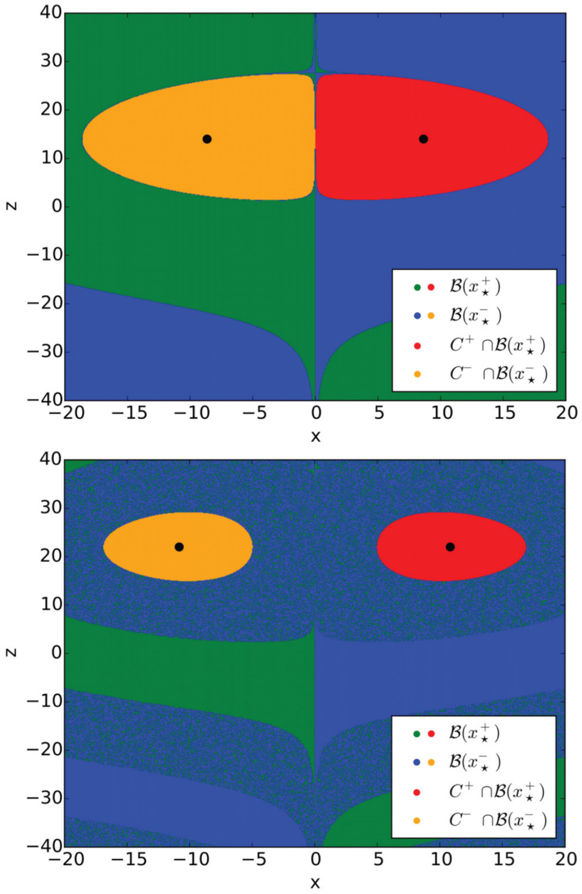


FIG. 2. Illustrations of the phase space structure of the L63 system defined by Eqs. (11)–(13). Both panels show cross sections, obtained by cutting along the plane containing the origin with normal (1,1,0), of the basins of attraction $\mathcal{B}(\mathbf{x}_*^\pm)$ and their intersections $C^\pm \cap \mathcal{B}$ with the sets $C^\pm$ for $r = 15$ (upper panel) and $r = 23$ (lower panel). In both panels, the blob-shaped region around each fixed point (black dots) is $C^\pm \cap \mathcal{B}$. Top panel: the total basin of a fixed point is composed of successive layers: it is given by $C^\pm \cap \mathcal{B}$ combined both with the region in the respective other half-space ($x > 0$ or $x < 0$) directly surrounding $C^\pm \cap \mathcal{B}$ and with the next layer of the same color in the fixed point's half plane. In the lower panel (coloring identical), the fractal structure of the basins, visible as intermingled green and blue sets, is apparent; it is associated with transient chaos. One observes that the fraction of the window covered by the sets $C^\pm$ shrinks as r increases. This illustrates the general behavior observed in the L63 system and quantified by CBS, namely that the volume fraction of the three-dimensional sampling region occupied by $C^\pm$ decreases continuously as r approaches r_3.

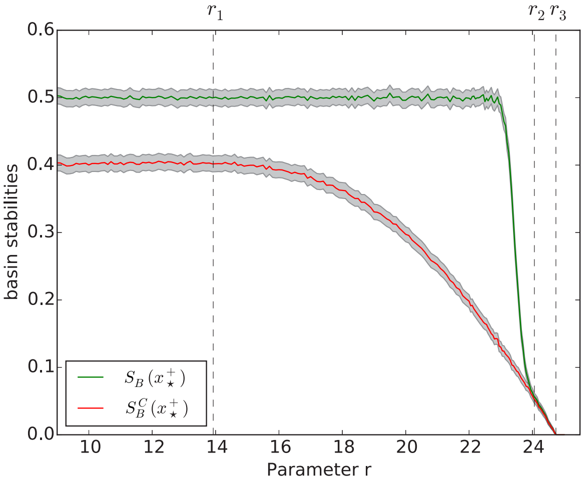


FIG. 3. BS [green (upper) line] and CBS [red (lower) line] of the positive fixed points in the L63 system as the parameter r is varied. From left to right, the vertical lines indicate the appearance of the chaotic set (r_1), the boundary crisis (r_2), and the stability loss of the fixed point (r_3). The difference between the two curves represents the fraction of the basin from where chaotic transients evolve. Note that BS exhibits a jump at r_2 which is blurred by the finite simulation time: the simulation stopped before all (increasingly long) transients had reached the fixed point. Because of the system's x-y symmetry the negative fixed point has the same BS and CBS. The gray envelopes represent $\pm 3\sigma$ according to Eq. (10).

originate. The magnitude of CBS reflects the opposite, i.e., the part of the basin from where trajectories fall into the fixed point without crossing H. For $9 \leqslant r \leqslant 14$, the two stability measures are constant but differ in their value. For $14 \leqslant r \leqslant 23$ the fraction of chaotic transients increases continuously (in agreement with Fig. 2), while the basin volume, i.e., BS, does not change. Between $r \approx 23$ and the bifurcation point at r_2, the basins of the fixed points collapse as the fraction of trajectories exceeding m crossings grows rapidly. In the limit $m \to \infty$, the basin size changes discontinuously at r_2: the fraction of the basin that corresponds to the chaotic transients at $r < r_2$ suddenly feeds the newly born attractor when $r > r_2$. In Fig. 3 the drop is not discontinuous due to finite simulation time: long transients do not reach the fixed points before the simulation is ended. From r_2 to r_3, both stability measures coincide because chaotic transients are absent due to the chaotic saddle now being attractive. For $r > r_3$, both BS and CBS vanish. In short, with BS alone we cannot anticipate the crisis at r_2. However, combining it with CBS, the emergence of the chaotic set can be observed by the increasing fraction of chaotic transients within the fixed points' basins. Although CBS does not predict the bifurcation at r_2, it does indicate the approaching crisis, while BS does not and thus it substantially complements the original BS by revealing additional information on the basin structure.

B. Anderies carbon cycle model

Anderies *et al.* [15] present a conceptual model of global carbon cycle dynamics in the Earth system formulated as a mass balance between three carbon stocks $\mathbf{x} = (c_a, c_t, c_m)$ which are nondimensional atmospheric c_a, terrestrial c_t, and marine stocks c_m, respectively. Formally, the model for the preindustrial case is given by

$$\dot{c_t} = P_{EN}(c_a,c_t) - H(c_t), \tag{15}$$

$$\dot{c_m} = D(c_a,c_m), \tag{16}$$

Reprinted paper with permission from A. Van Kan, J. Jegminat, J. F. Donges, and J. Kurths, Constrained basin stability for studying transient phenomena in dynamical systems, Phys. Rev. E, vol. 93, no. 4, pp. 1-7, 2016, doi: 10.1103/PhysRevE.93.042205. Copyright 2016 by the American Physical Society.

where total carbon in the system is conserved such that $c_a + c_t + c_m = 1$ and $c_a,c_t,c_m \geqslant 0$. The expressions describing the derivatives $\dot{c}_t,\dot{c}_m$ are defined as follows: a harvesting term $H(c_t) = \alpha c_t$, where α determines the human offtake of terrestrial carbon stocks, a diffusion term between atmosphere and ocean $D(c_a,c_m) = 0.05(c_a - c_m)$, and net ecosystem productivity $P_{EN}(c_a,c_t) = 2.5c_t(1-c_t/0.7)\{1.5c_a^{0.3}220T(c_a)^3 \exp[-7T(c_a)]-110T(c_a)^4 \exp[-5T(c_a)]\}$ with $T(c_a) = 0.8c_a + 0.2$.

It is found in [15] that any initial condition converges to one of two fixed points of interest: either a desirable state $\mathbf{x}_\star^d = ((c_a)_\star^d,(c_t)_\star^d,(c_m)_\star^d)$ with vegetation or an undesirable global desert state $\mathbf{x}_\star^{ud}$. At low values of $\alpha \in [0,0.6]$, the desirable state is attractive, while the undesirable state is repulsive. At $\alpha_{\text{crit}} \approx 0.4$ a transcritical bifurcation occurs and the fixed points reverse their stability. Anderies *et al.* [15] study their model in the context of planetary boundaries interacting with each other. In order to define a safe operating space, they suggest to classify trajectories by whether they return to a certain small ε ball around the desirable fixed point $\mathbf{x}_\star^d$ by a certain critical time t_{crit}. Translating this into our framework, we obtain the condition

$$C = \{\mathbf{x} \in \Gamma \mid \exists t < t_{\text{crit}} \text{ such that } |\phi^t(\mathbf{x}) - \mathbf{x}_\star^d| < \varepsilon\}, \quad (17)$$

where we choose $\varepsilon = 10^{-4}$. The constraint formulated in Eq. (17) is an integrated constraint since it depends on time. The motivation for this choice of constraint is that, although all trajectories converge to $\mathbf{x}_\star^d$ as $t \to \infty$ for $\alpha < \alpha_{\text{crit}}$ (since then $\mathbf{x}_\star^d$ is globally attractive), some trajectories pass very closely and slowly by $\mathbf{x}_\star^{ud}$. These trajectories would entail catastrophic consequences for life on the planet. Therefore, they are identified by whether they exceed a certain return time threshold. Thus we compute CBS of the desirable fixed point based on Eq. (17). We consider a perturbation density ρ describing a depletion of the terrestrial carbon stock c_t (e.g., by immense wildfires). The released carbon is fed into the atmospheric carbon stock c_a. We implement this scenario using a uniform perturbation density on a line in phase space: the terrestrial carbon stock is depleted to a value $c_t \in [0,(c_t)_\star^d)$, while the marine carbon stays constant, $c_m = (c_m)_\star^d$, and the atmospheric carbon increases according to the carbon conservation law $c_a = 1 - c_m - c_t$. We draw $N = 500$ initial conditions. Figure 4 shows the set C in two-dimensional phase space for two different values of α. The more detailed dependence of BS and CBS on α is shown in Fig. 5: BS is discontinuous at α_{crit} (within numerical accuracy), whereas CBS exhibits a smooth monotonic decay from 1 to 0 on the interval $\alpha \in [0.02,0.31]$. This reflects the fact that perturbations result in undesirable trajectories much more frequently as human carbon offtake increases until, at $\alpha \approx 0.31$, the return time for the considered perturbations always exceeds t_{crit}. Even though $\alpha < \alpha_{\text{crit}}$, none of the perturbed trajectories can avoid passing through a long quasiglobal desert state. On this parameter interval the desired state is unstable with respect to CBS but stable with respect to BS. We suggest that the former measure provides a more meaningful notion of Earth system resilience from an anthropocentric point of view: it measures the probability that perturbations decay within a predefined acceptable time horizon, while the latter only measures the probability of returning within any (possibly infinite) time horizon. Further, CBS captures the change in transient structure and therefore reveals a signal of the transition in the Anderies model already at values of α significantly smaller than α_{crit}. In contrast, BS is discontinuous (within numerical accuracy) at α_{crit} and does not exhibit any precursory phenomena. Figure 4 shows the set C defined above in Eq. (17) and $\mathcal{B}$, the basin of attraction of the desirable fixed point for two different values of α.

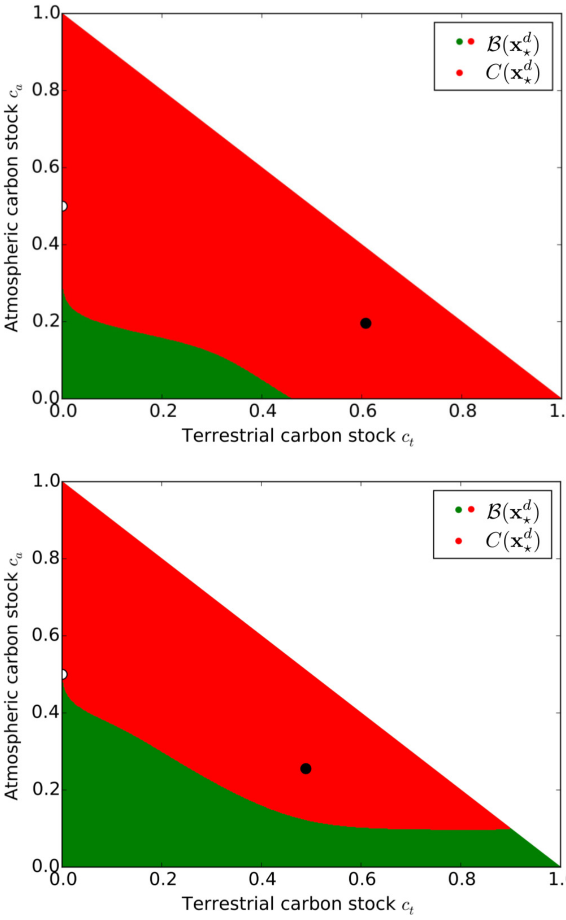


FIG. 4. Illustrations of the phase space structure of the Anderies system, Eqs. (15) and (16). Both panels show the (globally attracting) desirable fixed point (black dot) and the set C (red, upper part of triangle) and its complement $\mathcal{B}/C$ (green, lower part of triangle) for $\alpha = 0.15$ (upper panel) and $\alpha = 0.35$ (lower panel). The white dot at (0,0.5) is the desert state fixed point. One observes that the fraction of the two-dimensional finite phase space covered by the set C shrinks as α increases, in agreement with Fig. 5.

IV. DISCUSSION

We have defined CBS as a generalization of BS, thereby combining an asymptotic stability measure with information retrieved from transient behavior into a compact and intuitive measure. While BS is computed from an attractor's basin,

Reprinted paper with permission from A. Van Kan, J. Jegminat, J. F. Donges, and J. Kurths, Constrained basin stability for studying transient phenomena in dynamical systems, Phys. Rev. E, vol. 93, no. 4, pp. 1-7, 2016, doi: 10.1103/PhysRevE.93.042205. Copyright 2016 by the American Physical Society.

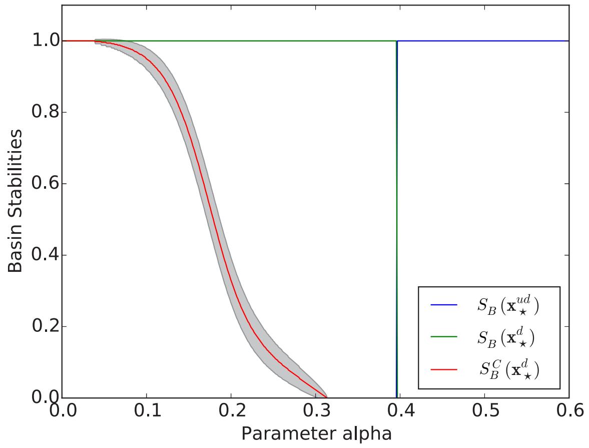


FIG. 5. BS (straight lines) and CBS (curved red line) of $\mathbf{x}_\star^d$ and $\mathbf{x}_\star^{ud}$ vs the human carbon offtake rate $\alpha \in [0,0.6]$ for $t_{\text{crit}} = 90$. BS of $\mathbf{x}_\star^d$ is represented by the green straight line ($\alpha \leqslant \alpha_{\text{crit}} \approx 0.4$) and BS of $\mathbf{x}_\star^{ud}$ by the blue straight line ($\alpha \leqslant \alpha_{\text{crit}}$). At low values of α, the desirable state is stable against any strength of perturbation while the desert state is unstable. For $\alpha > 0.03$, an increasing fraction of perturbation-induced trajectories takes longer than t_{crit} to return to the desirable fixed point until, at $\alpha \approx 0.32$, CBS vanishes. By contrast, BS of both fixed points exhibits a jump at α_{crit} and thus no precursory phenomena can be observed there. The gray envelope represents $\pm 3\sigma$ according to Eq. (10).

CBS is computed from a subset of the attractor's basin. The subset is defined by the transient behavior of trajectories originating from this subset. Thus CBS represents potentially very complicated transient behavior as an easy to interpret scalar quantity.

To underpin that a compact representation of transient behavior is highly relevant in applications, we have presented two examples using specific constraints on the transients. In the case of Rayleigh-Bénard dynamics in the scope of the Lorenz63 model we used the static constraint that the sense of rotation of convection rolls does not change. Here, CBS uncovers nonlinear precursory phenomena of a boundary crisis bifurcation. In the global carbon cycle model by Anderies *et al.* [15], we have studied the stability of the desirable state for a specific perturbation scenario under the premise that it can be restored within an acceptable time horizon. CBS reflects the fact that long return times to the attractor after a perturbation are not desirable. More generally, these applications demonstrate the following three main advantages of CBS over BS. (i) CBS provides useful information in the case of global attractors, while BS cannot be meaningfully applied (it is always equal to 1). (ii) Sudden changes in basin size are often preceded by a change in transient behavior. Extending linear notions of early warning signals for incipient bifurcations [6], CBS uncovers these nonlinear precursory phenomena in the case of the Lorenz63 model and helps anticipating the boundary crisis. (iii) CBS reflects the fact that certain perturbation-induced transients are often undesirable, e.g., long return times, thus allowing one to define highly relevant stability measures for a specific application.

The importance of BS lies in its applicability to a wide range of dynamical systems in various fields. The concept of CBS is even more general as it encompasses BS as a special case. However, to apply CBS, we must choose a specific constraint, such as a limit on the return time. This choice strongly depends on a specific application, revealing highly relevant information there but potentially not being as useful in other applications. By providing two examples of useful constraints and by defining CBS precisely, formally, and in close analogy to BS, we hope to facilitate the transfer of ideas between different applications and different generalizations of BS. For example, BS has been employed successfully to study power grid stability [2]. CBS could be used to develop more specific notions of stability, e.g., to impose that certain units recover quickly from megaoutages or to constrain the total energy loss on the way of recovery. Another example is ecology where BS has proven to be a useful concept and transients are important [18]. CBS could be used to quantify questions of how fast ecosystems recover or investigate potential early warning signals based on minimal abundances of certain species after transients. More generally, BS has successfully applied in resilience research [19,20] and we expect interesting results from further investigating the notion of constrained resilience based on constraints on transients. We expect that future work on CBS will yield a set of transient constraints that prove valuable across a wide range of different applications.

CBS can be used in both passive and active experimental settings. In the former, we have only limited or no control of the system, e.g., the Earth system. We start with some normative notion of undesirable transients as the time threshold in the Anderies example. Heitzig and Kittel [21] discuss desirability in relation to phase space topology. From a given notion of desirability, a constraint is derived. Then, CBS addresses the question of how stable the system is with respect to perturbation-induced transients given that only some of them are desirable. If a system parameter varies over time, CBS is capable of revealing a stability trend which can justify an action to reverse the parameter change. It remains an open problem how CBS can be inferred experimentally or from observational data if a satisfactory model of the system is not available. In principle, if long time series of some environmental parameter (e.g., forest cover on the Earth's surface) can be derived from measurements and if many natural perturbations can be observed in the data, such as volcanic eruptions, then these can be exploited to estimate CBS. In the active setting, we use CBS to foster our understanding of the system without needing a normative proposition. The constraint and the perturbation are chosen such that new information about the structure of the basin of attraction is revealed. This situation is analogous to the Lorenz63 example: restricting the number of flips between the two halves of phase space (i.e., the two convection senses), the basin of attraction can be subdivided according to the number of flips. Thus CBS helps to characterize a system that is subject to perturbations. In this active setting, it is easier to measure CBS: the system parameters can be chosen freely and the number of perturbations is not restricted by historic events.

The specific condition (14) is reminiscent partly of the concept of "viability" [22,23], although there are significant differences: in particular, in our case, there are no man-

Reprinted paper with permission from A. Van Kan, J. Jegminat, J. F. Donges, and J. Kurths, Constrained basin stability for studying transient phenomena in dynamical systems, Phys. Rev. E, vol. 93, no. 4, pp. 1-7, 2016, doi: 10.1103/PhysRevE.93.042205. Copyright 2016 by the American Physical Society.

agement options and, more importantly, we are considering deterministic dynamics while viability theory incorporates stochastic and more generally nondeterministic processes. Furthermore, Eq. (14) can easily be generalized by allowing for trajectories to "pierce through" H once or multiple times—these generalizations are not related to viability theory. Another concept that has certain features in common with the condition (14) is "survivability" [24]. There, too, a desirable region of phase space is designated as in the case of our choice of $C^{\pm}$. However, survivability does not incorporate the asymptotic nature of BS: it depends on the fraction of trajectories starting in a designated region of phase space spending a certain time exclusively in that region. In particular, it does not depend on which attractor trajectories converge to in the long-time limit. For these reasons, CBS is different from both viability theory and survivability and presents a broadly applicable concept for quantifying stability of an attractor with respect to a given not only small perturbation, uniting both the asymptotic features of BS and the transient features of survivability. In conclusion, CBS represents a general framework to quantify the stability of attractors with broad applicability in various fields with an interest in complex dynamical systems, ranging from physics and technology to sustainability science.

ACKNOWLEDGMENTS

The research reported in this article was developed within the working group "Complex systems perspectives on anthropogenic climate change" at the Natur- and Ingenieurwissenschaftliches Kolleg sponsored by the German National Academic Foundation (Studienstiftung des deutschen Volkes) and in the scope of the COPAN project on coevolutionary pathways at the Potsdam Institute for Climate Impact Research. J.F.D. thanks the Stordalen Foundation (via the Planetary Boundary Research Network PB.net) and the Earth League's EarthDoc program for financial support. We thank the anonymous referees for their useful comments which helped to improve the readability of this paper and we are grateful for insightful discussions with Jobst Heitzig and Boyan Beronov.

A.v.K. and J.J. contributed equally to the reported research.

[1] P. J. Menck, J. Heitzig, N. Marwan, and J. Kurths, Nat. Phys. **9**, 89 (2013).

[2] P. J. Menck, J. Heitzig, J. Kurths, and H.-J. Schellnhuber, Nat. Commun. **5**, 3969 (2014).

[3] P. Schultz, J. Heitzig, and J. Kurths, New J. Phys. **16**, 125001 (2014).

[4] T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley (eds.), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, Cambridge, UK, 2013).

[5] M. I. Rabinovich, P. Varona, A. I. Selverston, and H. D. Abarbanel, Rev. Mod. Phys. **78**, 1213 (2006).

[6] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. Van Nes, M. Rietkerk, and G. Sugihara, Nature (London) **461**, 53 (2009).

[7] W. Steffen, W. Broadgate, L. Deutsch, O. Gaffney, and C. Ludwig, Anthropocene Rev. **2**, 81 (2015).

[8] H.-J. Schellnhuber, in *Earth System Analysis: Integrating Science for Sustainability*, edited by H.-J. Schellnhuber and V. Wenzel (Springer, Berlin, 1998), pp. 3–195.

[9] H. J. Schellnhuber, Nature (London) **402**, C19 (1999).

[10] G. Petschel-Held, H.-J. Schellnhuber, T. Bruckner, F. L. Toth, and K. Hasselmann, Clim. Change **41**, 303 (1999).

[11] J. Rockström, W. Steffen, K. Noone, A. Persson, F. S. Chapin III, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber, B. Nykvist, C. A. de Wit, T. Hughes, S. van der Leeuw, H. Rodhe, S. Sorlin, P. K. Snyder, R. Costanza, U. Svedin, M. Falkenmark, L. Karlberg, R. W. Corell, V. J. Fabry, J. Hansen, B. Walker, D. Liverman, K. Richardson, P. Crutzen, and J. A. Foley, Nature (London) **461**, 472 (2009).

[12] W. Steffen, K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R. Carpenter, W. de Vries, C. A. de Wit *et al.*, Science **347**, 1259855 (2015).

[13] K. Raworth, Oxfam Policy Practice: Clim. Change Resil. **8**, 1 (2012), http://policy-practice.oxfam.org.uk/publications/a-safe-and-just-space-for-humanity-can-we-live-within-the-doughnut-210490.

[14] E. N. Lorenz, J. Atmos. Sci. **20**, 130 (1963).

[15] J. M. Anderies, S. Carpenter, W. Steffen, and J. Rockström, Environ. Res. Lett. **8**, 044048 (2013).

[16] J. A. Brander and M. S. Taylor, Am. Econ. Rev. **88**, 119 (1998).

[17] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (Westview Press, Boulder, 2014).

[18] L. Dai, K. S. Korolev, and J. Gore, Proc. Natl. Acad. Sci. U.S.A. **112**, 10056 (2015).

[19] C. S. Holling, Annu. Rev. Ecol. Syst. **4**, 1 (1973).

[20] C. Folke, S. Carpenter, B. Walker, M. Scheffer, T. Elmqvist, L. Gunderson, and C. Holling, Annu. Rev. Ecol., Evol., Syst. **35**, 557 (2004).

[21] J. Heitzig, T. Kittel, J. F. Donges, and N. Molkenthin, Earth Syst. Dynam. **7**, 21 (2016).

[22] J.-P. Aubin and P. Saint-Pierre, Decis. Making Risk Manag. Sustain. Sci. **2007**, 43 (2007).

[23] J.-P. Aubin, A. M. Bayen, and P. Saint-Pierre, *Viability Theory: New Directions* (Springer Science & Business Media, New York, 2011).

[24] F. Hellmann, P. Schultz, C. Grabow, J. Heitzig, and J. Kurths, arXiv:1506.01257.

Reprinted paper with permission from A. Van Kan, J. Jegminat, J. F. Donges, and J. Kurths, Constrained basin stability for studying transient phenomena in dynamical systems, Phys. Rev. E, vol. 93, no. 4, pp. 1-7, 2016, doi: 10.1103/PhysRevE.93.042205.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Survivability of Deterministic Dynamical Systems

Frank Hellmann[1,*], Paul Schultz[1,2,*], Carsten Grabow[1], Jobst Heitzig[1] & Jürgen Kurths[1,2,3,4]

Received: 21 March 2016
Accepted: 20 June 2016
Published: 13 July 2016

The notion of a part of phase space containing desired (or allowed) states of a dynamical system is important in a wide range of complex systems research. It has been called the safe operating space, the viability kernel or the sunny region. In this paper we define the notion of survivability: Given a random initial condition, what is the likelihood that the transient behaviour of a deterministic system does not leave a region of desirable states. We demonstrate the utility of this novel stability measure by considering models from climate science, neuronal networks and power grids. We also show that a semi-analytic lower bound for the survivability of linear systems allows a numerically very efficient survivability analysis in realistic models of power grids. Our numerical and semi-analytic work underlines that the type of stability measured by survivability is not captured by common asymptotic stability measures.

In almost all dynamical systems applicable to the real world, the stability of the system's stationary states (periodic orbits, chaotic attractors, etc.) is of key interest, because perturbations are never truly absent and initial data is never exactly determined. Nevertheless, the asymptotic stability of the system's attractors ensures that we can still extract sensible long-term information from our dynamical models.

Complementary to the notion of stability, one can analyse whether the system will remain in a desirable regime[1]. This becomes important when a model represents a system that we have influence on, either because we engineer its fundamental behaviour, or because there are management options. We often want to design the dynamics, or our interventions, such as to more easily keep the system in such a desired state. Note that the desirable region not necessarily contains a stationary state.

For the traditional notion of asymptotic stability against small perturbations, the key mathematical concept is the analysis of the linearised dynamics, in particular by means of the Lyapunov exponent or master stability function[2,3].

Real-world systems typically are multistable[4–6]. They have more than one stable attractor[7], and thus potentially exhibit a wide range of different asymptotic behaviours. The key question then becomes from which initial state which attractor is reached, i.e., to determine the basin of attraction of an attractor. Most work so far focused on the geometry of the basin of attraction[8] of desirable attractors, e.g. by finding Lyapunov functions[9–11].

A recent idea that has been found to be useful is to study a more elementary property, i.e. not which states go to an attractor, but just how many. This quantity, the volume of the basin of attraction of a given attractor, can then be interpreted as the stability of the system in the face of a random, non-small perturbation. It quantifies the probability that the typically non-linear response to such a perturbation will lead the system to a different, undesirable attractor. This probability is called the *basin stability* (S_B) of an attractor[12]. This is important for a number of applications where relevant system deviations are typically not small, for example in neuro science, system Earth or power grids.

One of the key appealing features of S_B is that, by studying just the volume rather than the shape of the basin of attraction, it becomes numerically tractable to analyse even very high-dimensional systems. It was also shown that the information revealed by the volume of the basin genuinely complements the information provided by the Lyapunov exponents of the system[12].

There are, however, two major drawbacks when estimating S_B. On the one hand, the measure relies on identifying the asymptotic behaviour of a system, which might be difficult to detect, typically requires prior knowledge

[1]Potsdam Institute for Climate Impact Research, P.O. Box 60 12 03, 14412 Potsdam, Germany. [2]Department of Physics, Humboldt University of Berlin, Newtonstr. 15, 12489 Berlin, Germany. [3]Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom. [4]Department of Control Theory, Nizhny Novgorod State University, Gagarin Avenue 23, 606950 Nizhny Novgorod, Russia. [*]These authors contributed equally to this work. Correspondence and requests for materials should be addressed to F.H. (email: hellmann@pik-potsdam.de) or P.S. (email: pschultz@pik-potsdam.de)

about the attractor's nature, and is only meaningful in multistable systems. On the other hand, a S_B estimation is insensitive to undesired transient behaviour of the system, i.e. if the trajectory visits an undesired part of the phase space where the system would take damage that is not modelled explicitly. To detect this type of dangerous transients, a new, complementary measure is required.

In this paper we introduce a new stability-related measure, the *survivability* $S(t)$ of a dynamical system. This is the fraction of initial system states (i.e. arising from an initial large perturbation) giving rise to evolutions that stay within a desirable regime up to a given time t. The set of these initial conditions is called *basin of survival*.

More formally, call the phase space of our system X, and a chosen desirable region $X^+ \subseteq X$. The finite-time basin of survival $X_t^S \subseteq X^+$ is defined as the set of initial conditions in X for which the entire trajectory over the interval $[0, t]$ lies in X^+. We choose a probability measure μ of initial conditions, reflecting our knowledge of the nature of perturbations we wish to study. Accordingly, the *finite-time survivability* is defined as

$$S_\mu(t) := \mu(X_t^S). \tag{1}$$

The total survivability then is the infinite-time limit of $S_\mu(t)$. This can naturally be decomposed into the probability that the initial perturbation is survived, and that the following trajectory stays save:

$$S_\mu(t) = \frac{\mu(X_t^S)}{\mu(X^+)}\mu(X^+) = S_{\mu^+}(t) \cdot \mu(X^+). \tag{2}$$

with $\mu^+(\cdot) := \mu(\,\cdot \cap X^+)/\mu(X^+)$. Now $\mu(X^+)$ does not depend on the dynamics but only on the desirable region and the perturbations, i.e. it is a constant for given X^+. The conditional survivability $S_{\mu^+}(t)$ captures the interplay of dynamics, desirable region and perturbations; it has a natural interpretation as the conditional probability of a system to survive random, large perturbations that do not kill it immediately.

Assuming a uniform distribution of perturbations, the measure μ is proportional to the volume Vol. The resulting conditional survivability is our main object of study in what follows. We will call this finite-time survivability of a dynamical system:

$$S(t) := S_{\mathrm{Vol}^+}(t) = \frac{\mathrm{Vol}(X_t^S)}{\mathrm{Vol}(X^+)}. \tag{3}$$

We are also interested in initial perturbations that only occur in a particular region of phase space. Thus, we want to study uniform perturbations in a subset $C \subset X$. The conditional survivability $S^C(t)$ can then simply be defined with respect to the measure $\mathrm{Vol}^C(\cdot) = \mathrm{Vol}(\cdot \cap C)/\mathrm{Vol}(C)$:

$$S^C(t) := S_{\mathrm{Vol}^C}(t) = \frac{\mathrm{Vol}(X_t^S \cap C)}{\mathrm{Vol}(C)} \quad . \tag{4}$$

An important example of such a conditionial survivability is the single node survivability for networked systems. There we condition on the phase space at a single node, thereby isolating the impact of local perturbations on the whole system. A mathematically precise discussion will follow in the power grid example in the results section and the supplementary information (SI).

To further illustrate this definition, consider a simple example: A penguin wishing to ski down a mountain X going the fastest route possible in Fig. 1. The system is multistable as the penguin might end up in the goal or the valley. However, if the penguin goes over the cliff it will almost certainly slide the rest of the way to the goal on its back. The state of the penguin is not explicitly modelled by our (potential) landscape. We take this into account by declaring the parts of the cliff our penguin can not ski safely as an undesirable region. Further, if the penguin wishes to continue skiing, the valley might or might not be undesirable as well. Depending on these choices, different starting points can be in the basin of survival. If the goal is the only desirable attractor, the basin of survival lies in its basin of attraction, but if the valley is OK, too, this is not the case, and the asymptotic structure plays no role.

As opposed to S_B or a linear(-ised) analysis based on Lyapunov exponents, the survivability is concerned not just with the asymptotic behaviour of the system, but depends strongly on the transient dynamics. As opposed to S_B it is applicable in unstable, mono-stable, or multistable, linear or non-linear systems.

The application of the survivability concept is especially appropriate when interventions happen at the same time scale as the system dynamics, or when entering an undesirable region is deadly.

A key insight is that evaluating survivability becomes amenable to Monte Carlo integration. This is due to focusing on the probability that the trajectory following a perturbation violates the boundary rather than trying to find the actual sets of phase space from which a trajectory survives. Hence, a survivability analysis, just as S_B, is applicable to very high-dimensional systems. In fact, the situation is more favourable than in the case of S_B, as the entire curve $S(t)$ can be evaluated at a computational cost not exceeding that of S_B, while potentially revealing much more information.

This sets survivability apart from formally similar approaches, e.g. in control theory[13,14]. Their precise relationship to survivability is discussed in detail in the *Methods* section.

For linear systems with a polyhedral desirable region, we derive a closed form lower bound on the *infinite-time survivability* $S_\infty := S(t \to \infty)$ as well as a semi-analytic, stronger bound that becomes exact in the case of vanishing dissipation. These bounds reveal that the survivability of linear systems depends strongly on the eigenvectors of the linear dynamics, rather than just the eigenvalues. The semi-analytic bound eliminates the need to simulate

Reproduced from: F. Hellmann et al., Survivability of Deterministic Dynamical Systems, Sci. Rep., vol. 6, no. 29654, pp. 1-12, 2016, doi: 10.1038/srep29654. Published under Creative Commons Attribution License 4.0 (CC BY).

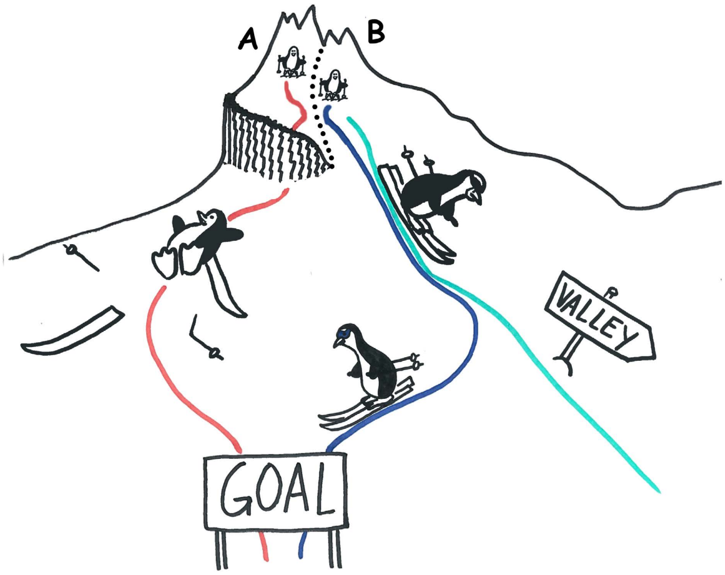


Figure 1. Survivability cartoon. A penguin can ski down the mountain starting anywhere on the slope. Starting at A the penguin will tumble over the cliffs, passing an undesirable state although ultimately reaching the goal. Starting at B the penguin will reach the goal standing on its feet. Starting even further to the right, it might end up in the valley, which might or might not be desirable.

the system trajectory opening survivability up to a wide range of applications for which numerically estimating the full dynamics is not feasible.

Results

To demonstrate the diverse applicability of our survivability concept we apply it to three paradigmatic model systems. A two-dimensional model of carbon stock dynamics, a system of integrate-and-fire neurons and a high-dimensional network model of the power grid.

These systems were chosen to cover a wide range of types of systems. The carbon cycle model has one or two attractors, depending on the parameter regime, and some transients are deadly. The neurons are mono-stable but exhibit transient chaos[15–19]. Finally the power grid model is high-dimensional, non-linear and multistable. However, the acceptable operating regime is close to a certain class of fixed points, thus the linearised behaviour near these fixed points is of great practical importance.

In all three systems there are externalities which are not or cannot be modelled explicitly. Namely, the influence of dramatic climate changes on society, external stimuli for a network of neurons and frequency control mechanisms in the power grid. We will see that survivability accurately captures the interplay of externalities with the intrinsic dynamics.

Carbon cycle model by Anderies *et al.*

We begin by applying survivability to a two-dimensional carbon cycle model from climate science which has been recently introduced[20]. This is a conceptual model with the aim to reproduce the non-linear dynamics of the carbon cycle in the Earth system. The boundaries of the *survival region* are closely related to the concept of planetary boundaries[21]. This system exhibits both the property that the undesirable states are deadly and that in some parameter regimes there is only a single stable attractor of the asymptotic dynamics.

The model equations for the atmospheric (c_a), marine (c_m) and terrestrial (c_t) carbon stocks are given by

$$\begin{aligned}\dot{c}_m &= \alpha_m(c_a - \beta c_m)\\ \dot{c}_t &= NEP(c_a, c_t) - \alpha c_t\\ c_a &= 1 - c_m - c_t\end{aligned} \tag{5}$$

where α_m denotes the atmosphere-ocean diffusion coefficient, β the carbon-solubility in sea water factor, α the human terrestrial carbon off-take rate and $NEP(c_a, c_t)$ the net ecosystem production, a complex non-linear relationship between the atmospheric and terrestrial carbon stocks (see Anderies *et al.*[20] for further details). Note that the total amount of carbon is kept constant, leaving us with the marine (c_m) and terrestrial (c_t) carbon stocks as independent variables.

Part of the phase space X of the model are states with virtually no terrestrial carbon, referred to as *desert states*. While the model can recover from such states and eventually reach high terrestrial carbon states again, entering a desert state would lead to the collapse of human civilisation and thus, tragically, our model would no longer be

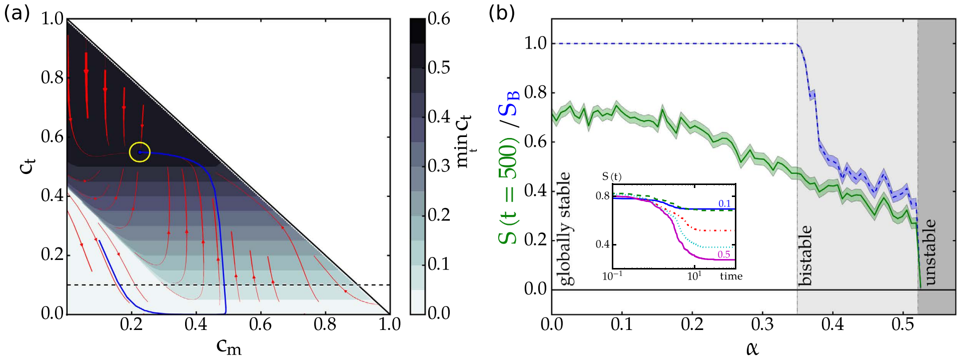


Figure 2. (**a**) Phase portrait of Anderies' model (Eq. 5, $\alpha = 0.1$). We choose initial terrestrial (c_t) and marine (c_m) carbon stocks, the colour scale then indicates the minimum of c_t over the whole time evolution commencing from a point. An example trajectory with a long excursion to the desert state ($c_t < m$) is plotted in blue and ends at the attractor which is circled in yellow, the stream plot indicates the vector field of the right-hand-side (cf. Eq. 5). The dashed black line indicates the value of the safety margin $m = 0.1$. (**b**) Bifurcations in the carbon cycle model. Basin stability (S_B, blue) and finite-time survivability ($S(t = 500)$, green) estimates for different values of the terrestrial human carbon off-take α. For the survivability estimation we assumed a safety margin $m = 0.1$. The shading around the curves indicates one standard error, the background colour indicates the different dynamical regimes. In the inset, we give survivability curves for five selected values of α, i.e. $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ from top to bottom as indicated.

valid after entering this regime. Hence, we define the set of desirable states X^+ as the complement of the desert states plus a safety margin m:

$$X^+ = \{(c_a, c_m, c_t) \in X: c_t > m\}. \tag{6}$$

The safety margin should at no time, during the transient or asymptotic behaviour, be crossed. The finite-time basin of survival, here introduced as X_t^S, is then given by

$$X_t^S = \left\{(c_a, c_m, c_t) \in X: \underset{0 \le t' \le t}{\forall} c_t(t') > m\right\}. \tag{7}$$

A phase plane analysis for this model is illustrated in Fig. 2(a). Of special importance here are those trajectories (exemplified by the blue trajectory in Fig. 2(a)) that first cross the safety margin, i.e. are not desirable due to the very low terrestrial carbon stocks c_t, but eventually will return to the desirable region X^+. These trajectories are counted for the S_B estimation, since they eventually approach the attractor, but are disregarded for the survivability, since they cross the safety margin during the transient period.

By varying the human carbon off-take α in Eq. 5, the system undergoes a bifurcation changing the number of attractors (around $\alpha = 0.35$) as illustrated in Fig. 2(b). The main picture shows the asymptotic survivability, the inset contains the survivability curves for different values of α. We see that the survivability drops to the asymptotic plateau at around the same time. Thus, if a trajectory eventually leaves the desirable regime, the time it takes until it does so is not strongly affected by α.

The bifurcation, which is known to be a saddle-node bifurcation[20], has a drastic impact on the S_B estimation, the survivability only changes marginally in this interval. On the other hand, the behaviour in the interval $\alpha \in [0; 0.35]$ shows how the S_B estimation becomes insensitive to system changes if the multistability is lost, i.e. if there is only a single attractor (in this case with non-zero c_t). The crucial question whether trajectories stay in a desired regime is thus not captured by the S_B measure, but can be answered with the survivability concept. Note that in this case and in what follows we estimate a finite-time survivability for the entire simulated time evolution of the system. Given that the asymptotic behaviour sets in earlier than the simulation ends, this is a good estimate for the infinite-time survivability.

It was argued[12] that S_B can also serve as a better early warning indicator of approaching tipping points than other measures. Here we see that a survivability estimation mirrors the trend in the system's behaviour, i.e. how the set of surviving states depends on system parameters, while S_B remains fixed at its plateau value. Hence, survivability can serve as a complementary, and in some scenarios better early warning sign than S_B.

Network of integrate-and-fire oscillators. In the case of transient chaos[15–18] there are long, interesting transients but potentially just a single global attractor. As an example, we consider a network of N integrate-and-fire neurons[22–25]. They exhibit long-term chaotic transients, but asymptotically have a global periodic attractor where the neurons are in a state of phase-synchronisation. Considering the synchronised state as undesirable, the integrate-and-fire neurons are an example of a system in which neither asymptotic nor basin stability are informative.

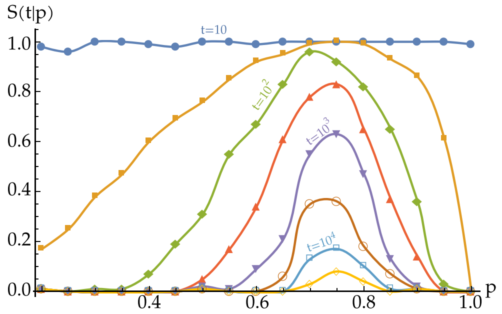


Figure 3. Survivability curves for networks of integrate-and-fire oscillators. Finite-time survivability $S(t|p)$ for given survival times t vs. the network parameter p. For each value of p we average over an ensemble of 100 network realisations, each with initial conditions drawn at random from the full state space.

Modelling external stimuli as essentially randomly resetting the phases of stimulated neurons, the survivability $S(t)$ here carries the interpretation of the probability that the system will not fall into a synchronised state in between stimuli, spaced apart at interval t. Such synchronised states model epileptic seizures and are thus undesired.

Concretely we study the convergence from arbitrary initial conditions to periodic orbit attractors, in which several synchronised groups of oscillators (clusters) coexist[26]. In the network every oscillator $j = 1\ldots N$ is connected to another oscillator $i \neq j$ by a directed link with probability p. A phase variable $\phi_j(t) \in [0, 1]$ specifies the state of each oscillator j at time t. The free dynamics of an oscillator j is given by

$$\dot{\phi}_j(t) = 1. \tag{8}$$

The oscillators interact on a directed graph by sending pulses when they reach the threshold $\phi_j = 1$. After a delay time τ this pulse induces a phase jump (indicated by differentiating the left and right limit of t as t^+ and t^-) in the receiving oscillator i:

$$\phi_i(t^+) := U^{-1}(U(\phi_i(t^-) - \varepsilon_{ij})) \tag{9}$$

for a potential U and coupling strength ε_{ij} (For more details cf. the SI).

The survivability $S(t|p)$ for a directed network of $N = 16$ pulse-coupled oscillators in dependence on the average connectivity p is illustrated in Fig. 3. For each value of p we create an ensemble of 100 network realisations. The randomly chosen initial phase vectors for each realisation are distributed uniformly in $[0, 1]^N$.

All different network realisations with their associated initial conditions eventually lead to a fully synchronous state. However, our concept of survivability reveals the highly non-linear, non-monotonic dependence on the network connectivity p. While the survivability of transient dynamic states is small for networks with low and high connectivity values p, it becomes very large for intermediate connectivities, even for only weakly diluted networks (Fig. 3). The finite-time survivability reveals a new, collective time scale that is much larger than the natural period, 1, of an individual oscillator and the delay time, τ, of the interactions.

These long, irregular transients are the main property of interest for the system, motivating their study in ref. 26. The dependence of the average lifetime of the transient chaotic trajectories on p was already studied ibidem. In this example, survivability reveals the same dynamical information as previous studies. Note that this is due to the specific choice of desirable region as the non-periodic parts of state space. Generally, there is no direct relationship between survivability and transient lengths, the fact that the desirable region can be chosen such that survivability reveals the quantity of interest for this system in a natural way speaks for its universality.

Survivability again is a natural and informative stability measure of this system, however, this time not against perturbations, but against getting trapped in an undesired corner of phase space.

Power grids. Power grids are subject to a variety of failures and perturbations and there are numerous studies concerning asymptotic stability analysis, e.g. refs 27 and 28, and recent approaches to an S_B assessment[29,30]. However, contrary to common model assumptions, the dynamical system does usually not evolve freely after a perturbation. If the system does not return to a stable operating state after a typical time span of a few seconds or if predefined thresholds are exceeded, control mechanisms that would require independent modelling are triggered.

The long-term behaviour and stability of the system is thus a question for control theory rather than just dynamics. Conversely, the transient dynamics, and the question whether there is a temporary amplification of perturbations, is critical to whether the control has to be activated at all, or the system is explicitly resilient to such perturbations. Hence, the power grid is an example where the undesirable region is deadly and management options operate at the system dynamics time scale.

The effective network model of the power grid[31,32] is the current standard baseline model for the frequency dynamics of power grids. It is known as the *swing equation* or the second-order Kuramoto model, and is used for short-term frequency stability studies in power grids. The various ways in which a power grid can be modelled using the swing equation are discussed in ref. 32 and limits to its applicability are discussed, for example, in refs 33 and 34.

The dynamical system modelling N generators' instantaneous phases ϕ_i and frequency deviations ω_i from the grid's rated frequency is given as

$$\begin{aligned}\dot{\phi}_i &= \omega_i\\ \dot{\omega}_i &= P_i - \alpha_i\omega_i - \sum_{j=1}^{N} K_{ij}\ \sin(\phi_i - \phi_j)\end{aligned} \tag{10}$$

with P_i being the net input power/consumption, α_i the electro-mechanical damping at node i and K_{ij} as the capacity of the link i – j. Here we choose $P_i = 1$ for net generators, $P_i = -1$ for net consumers, and a uniform distribution of $\alpha_i = \alpha = 0.1$. We choose the nonzero K_{ij} uniformly equal to 6, corresponding to an average transmission line length of about 200 km.

A stable operating state of the power grid is a fixed point of the dynamics with no frequency deviation, $(\phi^*, 0) := (\phi_1^*, \ldots, 0, \ldots)$. Conversely, limit cycle solutions (frequency oscillations) need to be prevented in order to avoid the tripping of generators. Frequency deviations are usually kept very small in large real power grids, with typical thresholds of ± 0.2 Hz[35] which corresponds to a phase velocity deviation of $|\omega| \approx 0.25$ in our units. Smaller island grids have considerably larger fluctuations. As an illustrative extreme case we will consider up to 20 times larger fluctuations. For S_B assessments, the reaction of the system to much larger deviations was also taken into account.

We will study the *single-node basin of survival*, i.e., the conditional basin of survival in the sense of Eq. 4, conditioned on initial perturbations that occur locally at a single node n, starting from a stable operating state. The space we wish to condition on is then the direct product of the stable operating state at all nodes except node n and the full state space of the node dynamics at n:

$$C_n = \{(\phi_1^*, \ldots \phi_n, \ldots, \phi_N^*, 0, \ldots, \omega_n, \ldots, 0) | \phi_n \in [0, 2\pi), \omega_n \in \mathbb{R}\}.$$

The desirable region being defined as $\forall i: |\omega_i| < 5$, which, as explained above, is chosen to mirror realistic constraints. Concretely, this means that we construct initial conditions by setting ϕ_i and ω_i to the value of the fixed point ϕ_i^* and 0, for all nodes other than the node n we are studying, and to a random phase in $[-\pi; \pi]$ as well as a random frequency deviation in $[-5; 5]$ for the node n. Then we simulate the system up to $t = 100$ and observe whether (and if, when) any of the frequency deviations ω_i leave the desirable region. In this way we sample 300 trajectories to estimate $S^n(t) := S^{C_n}(t)$. This leads to a standard error of less than 0.03 for $S^n(t) = 0.5$ in the worst case (see Methods section). We evaluate the survivability up to 100 in simulation time (18 s in real time), at which point a steady state has typically been established, and the asymptotic value of the survivability is reached.

While S_B captures the overall ability of the system to avoid permanent frequency oscillations, it does not directly capture the stability of the system against large perturbations. Instead, as discussed above, it is the ability of the system to keep perturbations under fixed frequency thresholds which is crucial. We will study this form of stability using both numerical simulations and the analytic approximations we have derived. The former will allow us to compare the survivability of the system to its S_B, the latter to assess the accuracy of our bounds.

We now turn to the question whether the semi-analytic bounds on the dynamics linearised around the fixed point can accurately mirror the single-node survivability $S^n(t = 100)$.

Defining $\phi := (\phi_1, \ldots, \phi_N)^T$, $\omega = (\omega_1, \ldots, \omega_N)^T$ and $\alpha := diag(\alpha_i)$, the linearised dynamics is given by

$$\begin{pmatrix}\dot{\phi}\\ \dot{\omega}\end{pmatrix} = \begin{pmatrix}0 & \mathbb{I}_N\\ L & -\alpha\end{pmatrix}\begin{pmatrix}\phi\\ \omega\end{pmatrix} \tag{11}$$

where the lower left block ($L = \partial\dot{\omega}_i/\partial\phi_j$) can be identified with the network's Laplacian matrix (at the fixed point $(\phi^*, 0)$) given by

$$L_{ij} = -\ \delta_{ij}\sum_{m=1}^{N} K_{im}\ \cos(\phi_i^* - \phi_m^*) + K_{ij}\ \cos(\phi_i^* - \phi_j^*) \tag{12}$$

The Jacobian has two real eigenvalues, $\lambda_1 = 0$ and $\lambda_2 = -\alpha$, corresponding to the eigenvectors $(\phi, \omega)_1 = (1, \ldots, 0, \ldots)$ and $(\phi, \omega)_2 = (-1/\alpha, \ldots, 1, \ldots)$. The first eigenvalue, λ_1 and the corresponding eigenvector show the linearised version of the rotational symmetry of the system under shifting all elements of ϕ by the same amount ϕ_s: $\phi_i \mapsto \phi_i + \phi_s$. The second corresponds to a homogeneous shift of all oscillator's frequencies, which does not affect the phase differences, and decays exponentially due to the damping term. The remaining part of the spectrum consists of $N - 1$ pairs of complex conjugated eigenvalues.

The basin of attraction in the conditional subspace C_n of this system is illustrated in Fig. 4(a). Concerning survivability, there is a subdivision in three different sets. The desirable region contains infinite- (central green region) and finite-time surviving states (yellow and red regions in the band). Trajectories commencing from the remaining states within the basin of attraction (blue region) eventually reach the attractor asymptotically. Note that there are also finite-time surviving states outside the basin of attraction (red region). A large part of the single-node basin of attraction is centred around the fixed point $(\phi^*, 0)$. Within this region we expect the linear approximation to provide a lot of information on the system.

Reproduced from: F. Hellmann et al., Survivability of Deterministic Dynamical Systems, Sci. Rep., vol. 6, no. 29654, pp. 1-12, 2016, doi: 10.1038/srep29654. Published under Creative Commons Attribution License 4.0 (CC BY).

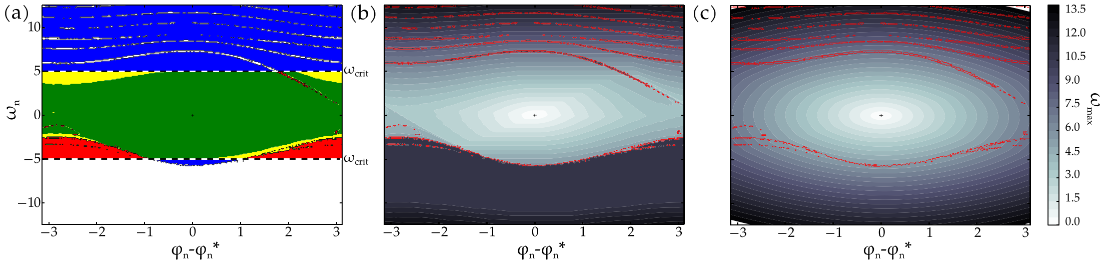


Figure 4. Single-node phase space of a consumer in the Scandinavian grid. (**a**) We plot the initial frequency deviation ω_n vs. the phase difference to the fixed point at node n, visualising the definition of the following areas using the simulation results from (**b**). The central green area resembles the infinite-time basin of survival, while the yellow and red areas contain finite-time surviving states. The union of the blue, yellow and green regions resembles the synchronous state's basin of attraction, while trajectories starting in the white or red regions approach different attractors. The frequency threshold is chosen as $\omega_{crit.} = \pm 5$ and initial conditions correspond to perturbations at a single consumer node of the network. (**b**) Simulated maximum frequency deviations ω_{max} along all dimensions, measured over the time evolution of the system for initial conditions that correspond to perturbations at node n of the network. For comparison with (**a**), we give the numerically estimated basin of attraction's boundaries in red. (**c**) Corresponding analytic upper bound for the maximum frequency deviation (cf. Eq. 16) for the linear approximation.

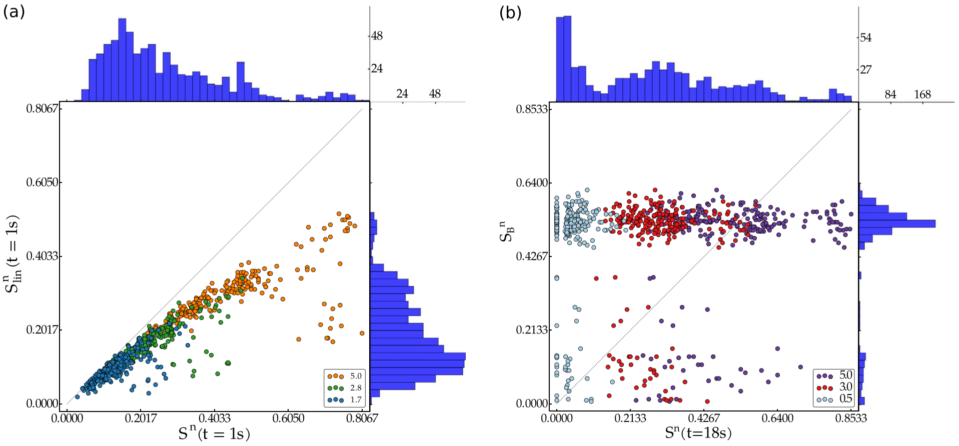


Figure 5. Simulated vs. approximated single-node survivability for the Scandinavian grid. (**a**) Scatter plot of the simulated $S^n(t)$ vs. approximated single-node survivability $S^n_{lin}(t)$ (cf. Eq. 16) estimated for all nodes in the Scandinavian power grid ($\omega_{crit.}$ is indicated in the legend). The corresponding distributions are given on the sides. (**b**) Single-node basin stability vs. single-node survivability for the Scandinavian grid. Scatter plot of the single-node basin stability S^n_B vs. single-node survivability $S^n(t = 100)$ ($\omega_{crit.}$ is indicated in the legend) estimated for all nodes in the Scandinavian power grid. The corresponding distributions are given on the sides. Note that we have chosen the initial region X_0 for single-node basin stability with $|\omega| < 100$, the same region as in ref. 29.

Regarding survivability, Fig. 4(b) shows that the frequency deviations inside the basin of attraction do indeed become large. The shape of the level lines of the frequency deviations corresponds to the basins of survival for different frequency constraints.

Figure 4(c) shows the bound for the frequency deviation of the linearised dynamics calculated according to Eq. 16. This shows a good qualitative agreement with the actually simulated frequency deviations as long as the deviations remain close to the fixed point, e.g. in the range of realistically allowed perturbations (see above). Still, the impact of the non-linearity (e.g. multistability is not captured) on the system becomes apparent, especially further away from the fixed point.

Indeed Fig. 5(a) shows that there is a high correlation between the lower bound of the survivability of the linear system $S^n_{lin}(t)$ calculated according to Eq. 16 (see *Methods section*) and the actual survivability $S^n(t)$ at the majority of nodes for realistic values of frequency deviations. What exactly gives rise to the outliers far below the diagonal will require further study. It is important to emphasise that the computational cost of calculating the bounds on the maximum frequency deviation for a sample of initial conditions is many orders of magnitude lower than the numerical estimate of the survivability via simulations of the actual time evolution. For a realistic network size of several hundred nodes, the approximate calculations can be performed on a laptop computer in

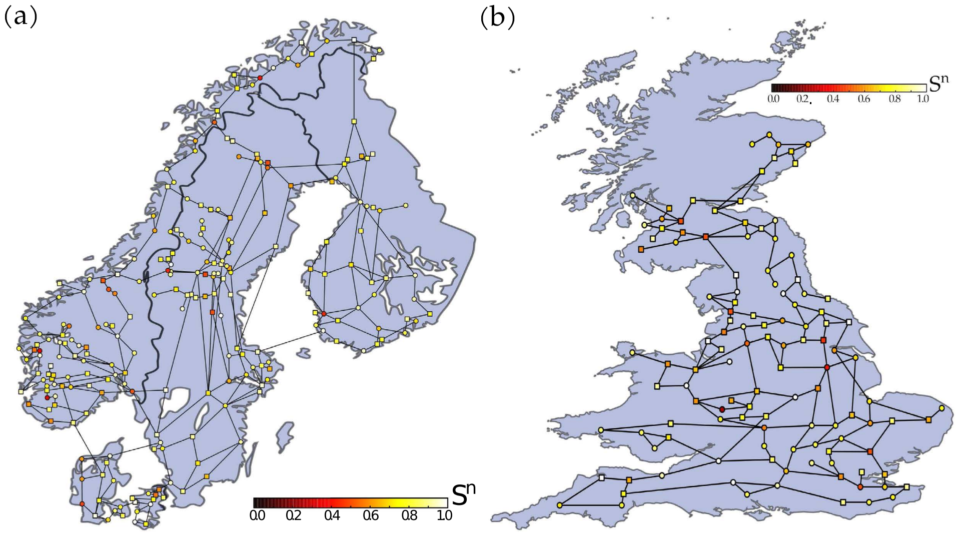


Figure 6. Scandinavian power grid. (**a**) The nodes' colouring indicates the respective single-node survivability estimate $S^n(t=1s)$ in the Scandinavian power grid. The frequency threshold is chosen as $\omega_{crit.}=\pm 10$. We randomly selected a dispatch scenario, circular nodes are net generators, squares are net consumers. The map of Scandinavia has been modified from https://commons.wikimedia.org/wiki/File:Scandinavia.svg, which is licensed under the Attribution-Share-Alike 3.0 Unported license. The license terms can be found on the following link: https://creativecommons.org/licenses/by-sa/3.0/. (**b**) UK power grid. Single-node survivability estimate $S^n(t=1s)$ of the UK power grid. Details analogous to (a). The map of Great Britain has been modified from https://commons.wikimedia.org/wiki/File:England,_Scotland_and_Wales_within_the_UK_and_Europe.svg, which is licensed under the Attribution-Share-Alike 3.0 Unported license. The license terms can be found on the following link: https://creativecommons.org/licenses/by-sa/3.0/.

less than a minute, whereas the numerical survivability estimation took several hours on 200 nodes of a computing cluster.

Figure 5(b) shows S_B^n as well as the single-node survivability of nodes in the Scandinavian power grid. We see that there is no significant correlation between the two quantities. This proves the point that the asymptotic behaviour of the system is not a strong indicator of the transient behaviour, at least in the case of power grids. The information we obtain from the survivability analysis is genuinely new information.

The Scandinavian power grid[29] consists of $N=236$ nodes and 320 links, corresponding to a mean degree of $\overline{k} = 2.7$. Hence, it has a sparse network topology with only a few neighbours per node on average, which is typical for power grids in general, independent from the number of nodes[36]. The same holds for our second data set, the UK high-voltage transmission grid, which consists of $N=120$ nodes and 165 links, corresponding to a mean degree of $\overline{k} = 2.8$.

In Fig. 6(a,b) we show the geographically embedded Scandinavian and UK power grid. The colour of each node corresponds to the single-node conditional survivability $S^n(t=1s)$. Different nodes exhibit starkly different survivability to perturbations. We find that at a threshold of $|\omega_{crit}|=10$, for both of these realistic power grid topologies, there are a few nodes that are particularly vulnerable to perturbations. This means a perturbation at these nodes is very likely to be amplified temporarily by the overall grid dynamics. What exactly leads to this vulnerability, and how to characterise it in terms of grid parameters and topology is a question for future work.

Finally, we also found that the survivability in this system asymptotes very quickly. Simulating just the first second of the power grid is typically sufficient, the so-called "first swing" following a disturbance mainly determines the overall frequency deviation.

Let us summarise the key points from applying survivability to power grids:

- For realistic small deviations, the upper bound applied to the linear approximation provides an excellent picture of the infinite-time basin of survival. The fact that the bulk of nodes shows a high correlation at large perturbations indicates that S^n can still be determined from the approximation in this case.
- For the given dynamics, the survivability very quickly reaches its asymptotic value. We expect this to be a fairly generic phenomenon if we are dealing with damped systems near a stable fixed point.
- Conditioning the survivability on regions of phase space with special meaning, like perturbations at a single node, allows us to reveal a large amount of non-obvious structural information on a networked system. Further work is needed to understand what gives rise to the revealed structure in realistic power grids.

Discussion

Survivability is a novel stability concept complementary to basin stability S_B and linear methods of asymptotic stability analysis. It applies to linear and non-linear systems, in the absence and presence of multi-stability. It focuses on transient rather than asymptotic behaviour, and incorporates exogenous information via assuming a desirable region for the system dynamics. Further, survivability can be estimated numerically at low computational costs, comparable to or even lower than for estimating S_B.

For linear systems we provide easy to evaluate analytic and semi-analytic expressions for lower bounds of the survivability, with a trade-off between the quality of the bound and numerical cost for evaluating the analytic expression. These reduce the need to simulate the system, yielding further dramatic improvements in computational cost.

The bounds we find demonstrate that the survivability depends crucially on the eigenvectors of the linear dynamics, rather than the eigenvalues (see discussion in the *Methods* section). It is an effective measure of the interaction between external constraints and the geometry of the dynamics in its phase space. The fact that the bound is tight exactly when the analysis of asymptotic stability using the eigenvalues of the linearised system fails shows that the survivability is genuinely complementary to eigenvalue-based stability concepts.

To explore this measure in practice, we analyse three conceptual examples.

Carbon Cycle. We observe that survivability accurately exhibits the presence of dangerous transient behaviour in the model, something that S_B can not detect. The almost monotonous decrease towards the first tipping point, opposed to the discontinuous S_B curve, shows the potential to derive an early warning scheme from an observation of these measures for certain kinds of bifurcations. Just as for S_B, the problem of evaluating the survivability from data remains a challenge for future work.

Neuronal Networks. Here, the transients do not arise from perturbations constructed as deviations around a desirable attractor, but they are randomly chosen from the whole compact phase space. Rather, the main interest lies on the transients themselves. Survivability reveals the same qualitative dependence of the dynamical behaviour on the underlying network topology as the average length of the transient[26]. Beyond that, considering $S(t)$ at fixed t as a function of the underlying topological parameters enables us to look in more detail into the relationship between function and structure of pulse-coupled oscillator networks. In contrast to the average length of the transients, the survivability also has a direct conceptual interpretation as the probability of the system remaining in the interesting transient regime. Thus it captures the appropriate notion of stability of transient chaos against the global attractor.

Power Grid. In this example we can see in detail the interplay between the semi-analytic bounds that we developed and the fully non-linear system. We demonstrate that survivability under realistic constraints captures information about the system not contained in the S_B estimate. We also demonstrate that the semi-analytic lower bounds, are strongly correlated with the simulations of the non-linear dynamics. Thus they contain much of the relevant information about the system. In strategic power grid development studies, this fact becomes particularly important as computational power is often at a considerable constraint, due to the need to simulate a wide range of divergent future scenarios of the energy transition. Dynamical properties outside of quasi-stationary calculations can only be taken into account if efficient estimators exist, since it is not feasible to run simulations. Thus our lower bounds, which eliminate the need for such simulations, potentially enable a more systematic way to investigate the impacts of the energy transition. In particular, the influence of changing topologies and different distributions of dynamical parameters on the dynamics of the power grid become computationally accessible. For the application to power grids, there are many more operational conditions on the system's behaviour that we do not consider here. While not all of them are as amenable to analytic considerations as the frequency deviation, we anticipate that it will still be possible to find cheap analytic boundaries for them. The reason that we could calculate the lower bounds so easily is that the phase space geometry is encoded in an efficient way in the eigenvectors. This aspect will carry over to many other, more complicated exogenous boundaries.

We thus have seen that the notion of survivability is general and powerful enough to capture the interplay between externalities and the intrinsic dynamics in three vastly different examples. In particular the last example demonstrated both the utility of single node survivability, revealing structural weaknesses and strengths of realistic power grid topologies, as well as of our semi-analytic bounds, reducing computational efforts dramatically.

The work presented here thus opens up a plethora of new avenues of research. On the theoretical side, the existence of a closed form lower bound on the survivability of a linear system opens the door to study the survivability as a function of the network topology and system parameters analytically, especially for the optimisation of these parameters to increase the system's survivability. The lower bounds presented here can certainly be improved by taking the more detailed geometry of the trajectories of the linear system into account. It will also be important to extend them to the types of bounds we have in more realistic power grid models.

Methods

Numerically estimating survivability.

One advantage shared by survivability and[12] is that they can be efficiently estimated by randomly sampling starting conditions. A trajectory either survives or not, therefore we can regard the sampling as a Bernoulli experiment with probability given by $S(t)$, hence the standard error (SE) of the probability estimator of a trial with N draws is simply

$$SE = \sqrt{\frac{S(t)(1 - S(t))}{N}}. \tag{13}$$

As a crucial consequence, the standard error of a survivability estimation does not depend on the dimensionality of the system. Further, the condition that a trajectory has left X^+ tends to be easier to evaluate in practice than whether the trajectory is asymptotically approaching a fixed point. Furthermore, in numerical simulations, an integration might be stopped once X^+ has been left.

Analytic results for linear systems. An important analytically tractable case is the total survivability S_∞ for a linear dynamic in $X = \mathbb{R}^N$, the Lebesgue measure $\mathrm{Vol}(X) = \int_X \mathrm{d}x^N$, and a polyhedral desirable region given by m linear conditions $y_k \cdot x(t) < 1$ for a set of vectors y_k, $k = 1 \ldots m$ in $\mathbb{R}^N$. In this case we can give a lower bound on $\mathrm{Vol}(X_\infty^S)$ that is easy to evaluate.

In this section we briefly give the results necessary for the applications in the results section on power grids. There we demonstrate that the semi-analytic bound captures the survivability of the system quite accurately in practical examples. In the SI we show detailed derivations, as well as further analytic results.

Consider a system of linear ordinary differential equations

$$\dot{x}(t) = Lx(t) \tag{14}$$

with $x \in X = \mathbb{R}^N$ and $L \in \mathbb{R}^{N\times N}$ with all eigenvalues having non-positive real parts. In general, L has a complex spectrum. The eigenvectors v_j of the complex eigenvalues are real or come in complex conjugate pairs, from which we pick one eigenvector each. We then define the $N \times N$ matrix $\mathbb{V}$ by stacking the eigenvectors, or their real and imaginary parts respectively, against each other as column vectors:

$$\mathbb{V} = \left[v_1, \ldots, \mathrm{Re}(v_j), \ldots, -\mathrm{Im}(v_j), \ldots\right]. \tag{15}$$

This allows us to translate initial conditions into the eigenvector basis by setting $c' = \mathbb{V}^{-1}x(0)$, and combining c_k' into complex numbers as appropriate $c_j = c_k' + ic_{k+n_c}'$, where n_c is the number of complex eigenvalues. Then the trajectory describes an exponential decay along the real eigenvectors and an inward spiral in the $\mathrm{Re}(v_j)$, $\mathrm{Im}(v_j)$ plane that is parametrised by c_j, and given by $\mathrm{Re}(\exp(\lambda_j t)c_j v_j)$. We then obtain an upper bound for the deviation of the trajectory starting at $x(0)$ in a direction y_k by maximising the contribution of each eigenvector separately.

Now, setting $y_{kj} := y_k \cdot v_j$ for v_j real, and $y_{kj} := |y_k \cdot v_j|$ for v_j complex, this leads to the estimate:

$$\max_{t\in[0;\infty[} |y_k \cdot x(t)| \le \sum_{j=1}^{n_0} y_{kj} c_j + \sum_{j=n_0+1}^{n_r} \max(0, y_{kj} c_j) + \sum_{j=n_r+1}^{n} y_{kj} \left|c_j\right| \tag{16}$$

where the first sum is over real eigenvectors corresponding to null eigenvalues, the second is over nonzero real eigenvectors and the last is over the complex eigenvectors.

Setting the right hand side of Eq. 16 smaller than 1 defines a region V_c in $\mathbb{R}^N$ spanned by the real and imaginary parts of the coefficients c_j. This region is mapped to the state space by $\mathbb{V}$ and thus its volume is related to the corresponding region in phase space by a determinant factor. As it is defined by a weaker inequality than X_∞^S it follows that

$$\mathrm{Vol}(X_\infty^S) \ge \sqrt{det \mathbb{V}\mathbb{V}^T}\, \mathrm{Vol}(V_c). \tag{17}$$

The inequalities Eq. 16 together with the matrix $\mathbb{V}$ can be used to efficiently estimate the total survivability as well as the conditional survivability. Remarkably, for systems with a purely imaginary spectrum, the bounds of Eqs 16 and 17 hold with equality.

In the SI we also derive a lower bound for $\mathrm{Vol}(V_c)$.

This lower bound demonstrates that for the survivability of a linear system, the eigenvectors play a crucial role. In fact, the eigenvalues do not enter the bound at all, except in terms of classifying the corresponding eigenvectors in separate classes. This demonstrates that the survivability captures substantially different information about the linear system than eigenvalue-based stability measures like relaxation time, or the master stability function.

Relationship to Similar Concepts. Survivability is related to a number of concepts in other fields, notably control theory. From this perspective it can be seen as a so far unstudied, simplifying case where a number of distinct concepts from various fields intersect. In this section we discuss a number of such concepts and their precise relationship to survivability.

Survivability is conceptually similar to the notion of finite time stability as studied for linear control systems[13,14]. There the focus is on finding a particular control scheme that will ensure that the resulting closed loop system stays within a particular region for some time, possibly in the presence of perturbations of the dynamical equations. From our perspective this can be seen as attempting to find systems with $S(t) = 1$. As the focus there is on perturbed dynamics in linear control systems, the actual overlap of methods is very small, in particular it is not possible to extend the methods to high-dimensional non-linear systems.

Another concept from control theory which is similar to the basin of survival is the viability kernel defined by Aubin *et al.* in the context of viability theory[37,38]. They introduce the notion of an environment K that contains all desirable states. Within the environment, there is the so-called viability kernel V[39,40] as the set of all initial conditions from which the system *can* stay within the environment. This basically is a more general version of our infinite-time basin of survival for non-deterministic systems or systems with multiple evolution paths and a management process. Consequently, $K\backslash V$ corresponds to the set of finite-time surviving states in deterministic systems. The viability kernel's volume is proposed as a measure of the degree of viability[38], in the limit of no control it thus reduces to our total survivability. However, we are not aware of this special case ever being considered in the context of viability theory. Whereas survivability measures the ability of the intrinsic dynamics to withstand perturbation, viability theory is concerned with the question of the power of control. Beyond this conceptual difference, evaluating survivability also requires very different technical methods, analytically as well as numerically. As far as we are aware, sampling based methods, which are efficient and natural for survivability, are impossible for viability. This is due to the fact that whether a particular point belongs to the viable set depends on the optimal control, which might not be known.

There are two concepts that share some formal similarity to survivability in the context of deterministic systems, transient times and open systems.

The study of transient life times[15,19,41,42] is only related to the survivability in the non-typical special case that the attractor (or a small epsilon environment around it) is the only undesirable region. In our example of integrate-and-fire neurons this is the case, but in the power grid there is no clear relationship between the strength of the transient (which might kill the system) and the return time to the attractor. In fact, there, the attractor we start from is in the desirable region. Transients life times are a special case, and not a typical one, of survivability. The latter is far more general, going beyond the focus on the length of transients and their distribution, and typically captures genuinely different information of the system (e.g. the linear analysis mainly depends on eigenvectors, not eigenvalues).

The theory of open systems, on the other hand, is generally concerned with ergodic systems. For leaky chaotic systems[43] the asymptotic behaviour of the survival probability is the key observable. At the formal level there is an analogy to our definitions, however, the total survivability, the size of the total phase space that leaks, is never considered as an observable in the literature. Indeed it is often the case that it is the whole phase space. Nor is the cumulative leakage ever interpreted as a stability measure or are efficient methods to estimate it for high-dimensional systems being discussed. In fact, as in the case of transient times, leaky systems can be seen as a special case of our discussion. Specifically it is the conditional survivability with the conditional space chosen as the space of surviving states X^S.

The closest analogy to our deterministic survivability is simply the survival analysis in the the context of stochastic systems. The concept of the so-called first hitting time and survival probability[44–46], which can be studied for the case of stochastic perturbations to deterministic systems by quasi-potentials[47–49], map directly to our work. The first hitting time t measures when a system is expected to first hit the forbidden region X^-. The cumulative of the probability of first hitting the undesirable region before t is then $1 - S(t)$. Our definitions given above can be seen as a deterministic version of these concepts. The role of stochasticity in the evolution is replaced by a probabilistic initial perturbation. Here similar sampling based methods are possible and necessary. The type of semi-analytic analysis we performed for the linear case would however be hard to duplicate. From this perspective what we have demonstrated is how to successfully apply methods and concepts from stochastic systems in the study of their deterministic counterparts.

The key insight in our work, as it is for S_B, is that restricting ourselves to probabilistic notions enables a considerably wider applicability of our analysis, as well as new numerical and analytic methods. Put differently, by asking not about the geometry of sets in phase space but merely about their volume, we can access high-dimensional non-linear systems that are out of reach for detailed geometric analysis. The challenge then lies in defining interesting sets that capture concepts of interest. As such we take it as a confirmation for the wide interest of the specific sets that survivability is based on, that it occurs a the intersection of a number of well studied concepts.

References

1. Heitzig, J., Kittel, T., Donges, J. F. & Molkenthin, N. Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the Earth system. *Earth System Dynamics* **7,** 21–50 (2016).
2. Nishikawa, T. & Motter, A. E. Synchronization is optimal in nondiagonalizable networks. *Physical Review E* **73,** 065106 (2006).
3. Pecora, L. M. & Carroll, T. L. Master Stability Functions for Synchronized Coupled Systems. *Physical Review Letters* **80,** 2109–2112 (1998).
4. Feudel, U. & Grebogi, C. Why are chaotic attractors rare in multistable systems? *Physical Review Letters* **91,** 134102 (2003).
5. Pisarchik, A. N. & Feudel, U. Control of multistability. *Physics Reports* **540,** 167–218 (2014).
6. Shrimali, M. D., Prasad, A., Ramaswamy, R. & Feudel, U. The Nature of Attractor Basins in Multistable Systems. *International Journal of Bifurcation and Chaos* **18,** 1675–1688 (2008).
7. Milnor, J. On the concept of attractor. *Communications in Mathematical Physics* **99,** 177–195 (1985).
8. McDonald, S. W., Grebogi, C., Ott, E. & Yorke, J. A. Factal Basin Boundaries. *Physica 17D* **17,** 125–153 (1985).
9. Belykh, V. N., Belykh, I. V. & Hasler, M. Connection graph stability method for synchronized coupled chaotic systems. *Physica D: Nonlinear Phenomena* **195,** 159–187 (2004).
10. Chiang, H.-D. *Direct Methods for Stability Analysis of Electric Power Systems* (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2010).
11. Zwillinger, D. *Handbook of differential equations* (Academic Press, 1997).
12. Menck, P. J., Heitzig, J., Marwan, N. & Kurths, J. How basin stability complements the linear-stability paradigm. *Nature Physics* **9,** 89–92 (2013).
13. Amato, F., Ariola, M., Cosentino, C., Abdallah, C. & Dorato, P. Necessary and sufficient conditions for finite-time stability of linear systems. In *Proceedings of the American Control Conference 2003*, vol. **5,** 4452–4456 (2003).
14. Amato, F., Ariola, M. & Dorato, P. Finite-time control of linear systems subject to parametric uncertainties and disturbances. *Automatica* **37,** 1459–1463 (2001).
15. Houghton, S., Knobloch, E., Tobias, S. & Proctor, M. Transient spatio-temporal chaos in the complex Ginzburg-Landau equation on long domains. *Physics Letters A* **374,** 2030–2034 (2010).
16. Tél, T. Transient Chaos. In Hao, B.-l. (ed.) *Directions in Chaos*, vol. 3, 149–221 (World Scientific, Singapore, 1990).
17. Tél, T. Transient chaos: a type of metastable state. In *Statphys*, vol. 19, 346–362 (1996).
18. Wolfrum, M. & Omel'chenko, O. E. Chimera states are chaotic transients. *Physical Review E* **84,** 2–5 (2011).
19. Lai, Y.-C. & Tél, T. *Transient Chaos*, vol. 173 of *Applied Mathematical Sciences* (Springer New York, New York, NY, 2011).
20. Anderies, J. M., Carpenter, S. R., Steffen, W. & Rockström, J. The topology of non-linear global carbon dynamics: from tipping points to planetary boundaries. *Environmental Research Letters* **8,** 044048 (2013).
21. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461,** 472–475 (2009).

22. Ernst, U., Pawelzik, K. & Geisel, T. Synchronization induced by temporal delays in pulse-coupled oscillators. *Physical Review Letters* **74,** 1570 (1995).
23. Jahnke, S., Memmesheimer, R.-M. & Timme, M. Stable irregular dynamics in complex neural networks. *Physical Review Letters* **100,** 048102 (2008).
24. Mirollo, R. & Strogatz, S. Synchronization of pulse-coupled biological oscillators. *Siam Journal on Applied Mathematics* **50,** 366 (1990).
25. Winfree, A. T. *The geometry of biological time* (Springer, New York, 2001).
26. Zumdieck, A., Timme, M., Geisel, T. & Wolf, F. Long Chaotic Transients in Complex Networks. *Physical Review Letters* **93,** 244103–244104 (2004).
27. Dörfler, F., Chertkov, M. & Bullo, F. Synchronization in complex oscillator networks and smart grids. *PNAS* **110,** 2005–2010 (2013).
28. Motter, A. E., Myers, S. A., Anghel, M. & Nishikawa, T. Spontaneous synchrony in power-grid networks. *Nature Physics* **9,** 191–197 (2013).
29. Menck, P. J., Heitzig, J., Kurths, J. & Schellnhuber, H. J. How dead ends undermine power grid stability. *Nature Communications* **5,** 1–8 (2014).
30. Schultz, P., Heitzig, J. & Kurths, J. Detours around Basin Stability in Power Networks. *New Journal of Physics* **16,** 125001 (2014).
31. Filatrella, G., Nielsen, A. H. & Pedersen, N. F. Analysis of a power grid using a Kuramoto-like model. *The European Physical Journal B* **61,** 485–491 (2008).
32. Nishikawa, T. & Motter, A. E. Comparative analysis of existing models for power-grid synchronization. *New Journal of Physics* **17,** 15012 (2015).
33. Weckesser, T., Johannsson, H. & Ostergaard, J. Impact of model detail of synchronous machines on real-time transient stability assessment. In *2013 IREP Symposium Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid*, 1–9 (IEEE, 2013).
34. Auer, S., Kleis, K., Schultz, P., Kurths, J. & Hellmann, F. The impact of model detail on power grid resilience measures. *The European Physical Journal Special Topics* **225,** 609–625 (2016).
35. UCTE. Continental Europe Operation Handbook, Glossary. *Tech. Rep.* (2004).
36. Schultz, P., Heitzig, J. & Kurths, J. A random growth model for power grids and other spatially embedded infrastructure networks. *The European Physical Journal Special Topics* **223,** 1–18 (2014).
37. Aubin, J.-P. Viability Kernels and Capture Basins of Sets Under Differential Inclusions. *SIAM Journal on Control and Optimization* **40,** 853–881 (2001).
38. Aubin, J.-P., Bayen, A. & Saint-Pierre, P. *Viability Theory. New Directions* (Springer Science & Business Media, 2011).
39. Bonneuil, N. Computing the viability kernel in large state dimension. *Journal of Mathematical Analysis and Applications* **323,** 1444–1454 (2006).
40. Maidens, J. N., Kaynama, S., Mitchell, I. M., Oishi, M. M. K. & Dumont, G. A. Lagrangian methods for approximating the viability kernel in high-dimensional systems. *Automatica* **49,** 2017–2029 (2013).
41. Politi, A. & Torcini, A. Stable Chaos. In *et al.*, M. (ed.) *Nonlinear Dynamics and Chaos: Advances and Perspectives*, 103–129 (Springer-Verlag, Berlin Heidelberg, 2010).
42. Rosin, D. P., Rontani, D., Haynes, N. D., Schöll, E. & Gauthier, D. J. Transient scaling and resurgence of chimera states in coupled Boolean phase oscillators. *Physical Review E* **1,** 5 (2014).
43. Altmann, E. G., Portela, J. S. E. & Tél, T. Leaking chaotic systems. *Reviews of Modern Physics* **85,** 869–918 (2013).
44. Anishchenko, V., Astakhov, V., Neiman, A., Vadivasova, T. & Schimansky-Geier, L. *Dynamics of Chaotic and Stochastic Systems* (Springer, Berlin, 2006).
45. Ebeling, W. & Sokolov, I. M. *Statistical Thermodynamics and Stochastic Theory of Nonequilibrium Systems* (World Scientific, Singapore, 2005).
46. Redner, S. *A Guide to First-Passage Processes* (Cambridge University Press, 2007).
47. Freidlin, M. I. & Wentzell, A. D. *Random perturbations of dynamical systems*, vol. 260 (Springer Science & Business Media, 2012).
48. Graham, R. & Tél, T. Existence of a potential for dissipative dynamical systems. *Physical Review Letters* **52,** 9 (1984).
49. Kraut, S. & Feudel, U. Enhancement of noise-induced escape through the existence of a chaotic saddle. *Physical Review E* **67,** 015204 (2003).

Acknowledgements

We acknowledge gratefully the support of BMBF, CoNDyNet, FK. 03SF0472A, of the EIT Climate-KIC project SWIPO and Nora Molkenthin for illustrating our illustration of the concept of survivability using penguins. We thank Martin Rohden for providing us with the UK high-voltage transmission grid topology and Yang Tang for very useful discussions. The publication of this article was funded by the Open Access Fund of the Leibniz Association.

Author Contributions

F.H. and P.S. designed the study; F.H., P.S. and C.G. prepared the data; F.H. and P.S. carried out the analysis and prepared the manuscript. All authors discussed the results and contributed to editing the manuscript. J.H. and J.K. supervised the study.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Hellmann, F. *et al.* Survivability of Deterministic Dynamical Systems. *Sci. Rep.* **6**, 29654; doi: 10.1038/srep29654 (2016).

Reproduced from: F. Hellmann et al., Survivability of Deterministic Dynamical Systems, Sci. Rep., vol. 6, no. 29654, pp. 1-12, 2016, doi: 10.1038/srep29654. Published under Creative Commons Attribution License 4.0 (CC BY).

IOP Publishing New J. Phys. 19 (2017) 083005 https://doi.org/10.1088/1367-2630/aa7b61

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics
Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
16 November 2016

REVISED
20 June 2017

ACCEPTED FOR PUBLICATION
23 June 2017

PUBLISHED
7 August 2017

PAPER

Timing of transients: quantifying reaching times and transient behavior in complex systems

Tim Kittel[1,2], Jobst Heitzig[1], Kevin Webster[1] and Jürgen Kurths[1,2,3]

[1] Potsdam Institute for Climate Impact Research, Telegrafenberg A31—(PO) Box 60 12 03, D-14412 Potsdam, Germany
[2] Institut für Physik, Humboldt-Universität zu Berlin, Newtonstraße 15, D-12489 Berlin, Germany
[3] Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom

E-mail: Tim.Kittel@pik-potsdam.de

Keywords: early-warning signals, complex systems, nonlinear dynamics, ordinary differential equations, stability against shocks

Supplementary material for this article is available online

Abstract

In dynamical systems, one may ask how long it takes for a trajectory to reach the attractor, i.e. how long it spends in the transient phase. Although for a single trajectory the mathematically precise answer may be infinity, it still makes sense to compare different trajectories and quantify which of them approaches the attractor earlier. In this article, we categorize several problems of quantifying such transient times. To treat them, we propose two metrics, area under distance curve and regularized reaching time, that capture two complementary aspects of transient dynamics. The first, area under distance curve, is the distance of the trajectory to the attractor integrated over time. It measures which trajectories are 'reluctant', i.e. stay distant from the attractor for long, or 'eager' to approach it right away. Regularized reaching time, on the other hand, quantifies the additional time (positive or negative) that a trajectory starting at a chosen initial condition needs to approach the attractor as compared to some reference trajectory. A positive or negative value means that it approaches the attractor by this much 'earlier' or 'later' than the reference, respectively. We demonstrated their substantial potential for application with multiple paradigmatic examples uncovering new features.

1. Introduction

In complex dynamical systems, the importance of a trajectory's transient, i.e. the part of the trajectory distant from the attractor, has been identified in physics research as well as in various other fields. Different phenomena during the process of magnetization for various materials, in particular the domain growth, have been studied extensively [1–3]. In laser physics, it was possible to derive analytical results matching the transient phases of different lasers [4, 5]. In kinetic theory, there has been research on non-equilibrium approaches for more than a century by now [6]. Other modern areas of statistical physics have emphasized the importance of transient dynamics, too, e.g. in social systems [7] and transient phases between jam and free-flow phases in vehicular traffic [8].

Even outside of the direct field of physics, but still within the scope of complex dynamical systems, a focus on transient dynamics has been developed recently. Hastings [9] made a call for more transient analysis of ecological models. An example of this was given by van Geest [10], describing macrophyte-dominated states of lakes as non-equilibrium states. In medicine and biology, epilepsy is seen as a transient phenomenon and much work has been done [11, 12]. In economics, a transient analysis complementing the asymptotic analysis proved to be fruitful, particular supporting with the stability analysis and understanding how to reach the equilibria [13]. Climate change is often seen as a transition to a new situation, i.e. a transient change to a new attractor [14–16]. Closely related, discussions in sustainability sciences are on transient dynamics because they refer to transformations from and to sustainability. Important key topics are the Anthropocene [17–19], particular the great acceleration [20], and planetary boundaries [21, 22].

Reproduced from: T. Kittel et al., Timing of transients: Quantifying reaching times and transient behavior in complex systems, New J. Phys., vol. 19, no. 8, 2017, doi: 10.1088/1367-2630/aa7b61. Published under Creative Commons Attribution License 3.0 (CC BY).

An important emphasis on *long transients* has been made in [10, 15, 23]. With this term, they refer to trajectories where the relevant and observable phenomena/states, e.g. macrophyte-covered lakes or desert states of the Earth system, are away from the actual attractor, but in the transient phase where a trajectory may stay for a substantial amount of time.

Hastings [9] stressed the importance of different time scales and pointed out how the transient dynamics can be very different and much more interesting than the asymptotic behavior. In addition, he explained how saddles play a central role by inducing long transients. This has been demonstrated in a study by Anderies *et al* [15] in the context of sustainability science. The 'interacting planetary boundary' [21, 22] has been defined by whether states take 'long' to the attractor or not. This idea leads precisely to the main question for this article 'How can we properly quantify the time to reach a system's attractor?', i.e. associate meaningful numbers with it.

This study is meant as a methodological step in direction for applications in real-world systems. So in the following, we focus on being able to do numerical estimations while analytical results are only given to understand general properties.

Often, a trajectory is divided arbitrarily in a transient part and the asymptotics close to the attractor. So we split the main question into two sub-questions: (a) 'What are the problems of these current/intuitive methods to quantify transient time?' and (b) 'How can we mend them?'.

To answer the first question, we work out four essential problems one is confronted with: (I) *infinite reaching time*: the attractor is not reached in finite time for a large class of physically relevant systems; (II) *physical interpretation*: it is unclear how to define precisely 'when the transient is over', so it is ambiguous where to divide between the transient and the asymptotics; (III) *discontinuities*: when having parameter dependence, small changes in the parameter often induce a large (noncontinuous) effect on the measured quantity; and (IV) *non-invariance*: the results depend on the choice of coordinates. Problem (IV) is particularly important, as a result should be a property of the dynamical system and thus independent of the choice of coordinates, i.e. invariant (or correctly transforming) under smooth transformations of the state space (see 'smoothly equivalent' in [24]).

Then, we approach question (b) by formulating two metrics, *area under distance curve* (D) and *regularized reaching time* (T_{RR}). The first one is the integral over the distance to the attractor along the trajectory, and has a physical dimension of time times distance. It measures which trajectories are *reluctant*, i.e. stay distant from the attractor for long, or *eager*, i.e. approach it right away. The second one, T_{RR}, is defined by the difference between the reaching times for the trajectory of interest and a reference trajectory. Thus, it takes a different point of view, actually measuring a time. The idea is that even though the actual reaching times are infinite (problem (I)), their difference is typically finite. So, we can compare trajectories approaching the attractor and define the notions *earlier* and *later*.

We chose four examples to illustrate different features of these metrics. We first use a linear system to understand how the metrics act generally and to observe the divergence of T_{RR} on the strong stable manifold particularly. Also, due to the system's simplicity, analytical solutions are possible. We then use a global carbon cycle model [15] and a model of a generator in the power grid [25] to apply the ideas to some first real world systems. Our final example, the chaotic Rössler oscillator, demonstrates that one can apply these methods to more complex attractors also, in this case a chaotic one. The chosen examples are rather well-understood. So they are good testing cases for the metrics, while their complexity still needs numerical approaches for a proper quantification of reaching times.

Finally, a detailed discussion on how far the two metrics solve the aforementioned problems is given, followed by a summary and an outline of future research.

The remainder of this article is structured as follows. After stating the four essential problems of reaching time definitions in section 2, we illustrate them with a small example. Then, we present the two metrics in section 3 and apply them to examples in section 4. Next, we give a detailed discussion on how far the metrics solve the essential problems in section 5. Finally, we close with a summary and an outlook. Additional information that can be found the supplemental material is available online at stacks.iop.org/NJP/19/083005/mmedia referenced within the article with the prefix 'Suppl. Mat.'.

Assumptions and notations. We assume a general, deterministic and autonomous dynamic system given by the differential equation

$$\dot{x} = f(x) \qquad x \in X \tag{1}$$

with an n-dimensional state space $X = \mathbb{R}^n$ and the right-hand side (rhs) $f(x)$. Usually, we use x to denote an arbitrary state $x \in X$, and refer to specific/fixed states with letters as superscripts, e.g. x^a, x^b, and x^{ref}. The components of a state are written with subscripts, so $x = (x_0, x_1, \ldots, x_{n-1})^\top$ and $x^a = (x_0^a, x_1^a, \ldots, x_{n-1}^a)^\top$. The words 'point' and 'state' are used synonymously for the elements of X. We assume the system (1) to have at least one attractor $\mathcal{A} \subseteq X$ with a basin of attraction $\mathcal{B}_{\mathcal{A}} \subseteq X$. In case the system has more than one attractor, the analysis should be applied to the attractors of interest separately.

For convenience, we will make heavy use of the time-evolution operator φ where $\varphi(t,\, x)$ is the state after starting at some point x and letting the system evolve for some time $t \geqslant 0$. Hence

$$\varphi(0,\, x) = x \quad \text{and} \quad \frac{\partial\varphi}{\partial t}(t,\, x) = f(\varphi(t,\, x)). \tag{2}$$

When we speak of 'quantifying the transient time' we mean to find a function $X \longrightarrow \mathbb{R}$, a 'metric', that gives a reasonable number for the time a trajectory spent in transient phase for each initial condition $x \in \mathcal{B}_{\mathcal{A}}$.

Additionally, within the article we assume the asymptotics of the system to be understood as we want to focus on the transient only.

2. The problems of reaching time definitions

In this section, we introduce four essential problems. They need to be addressed when aiming to quantify the transient time to reach an attractor $\mathcal{A}$ in a system of type (1). Then, we illustrate them with an example model.

(I) Infinite reaching time. A basic property of a large class of complex systems is that trajectories reach the attractor in infinite time only. That includes even steady states or limit cycles and most systems of ordinary differential equations with smooth rhs functions. This is the fundamental problem why the analysis made in this article is necessary.

(II) Physical interpretation. It is far from being obvious what the terms 'close to the attractor' or 'when the transient is over' means. Often, this is tackled by using some arbitrary threshold ϵ to define what is a 'small distance' to $\mathcal{A}$. But because of problem (I), the time to reach this ϵ-neighborhood typically diverges for $\epsilon \longrightarrow 0$. So the result depends strongly on the value of ϵ. Note that the focus of this article is to *quantify* the transient time to reach the attractor. So we want to associate meaningful numbers and need to treat this problem.

(III) Discontinuities. When defining a metric to quantify the transient time to reach $\mathcal{A}$ using some parameters e.g. ϵ, the result might depend discontinuously on the parameter. Usually, we want results to change smoothly and, if possible, weakly to changes of the parameter. If there is a discontinuous dependence, then we would expect there to be a corresponding specific property of the system that introduces this behavior.

(IV) Non-invariance. Our focus is on real-world systems. So the transient time should be a general property of the system, and not dependent on the chosen variables or coordinates to represent it. These coordinates correspond to a point of view on the system only. In other terms, invariance under change of coordinates should be given.

Example. While the aforementioned problems are of general nature, we illustrate them next using the example system

$$\dot{x}_0 = 1 - \frac{x_1}{2} - bx_0, \qquad \dot{x}_1 = 2(a - x_0^2), \qquad a = 2, \qquad b = 0.3. \tag{3}$$

It has a stable focus $x^{\mathrm{s}} = (-\sqrt{a},\, 2(1 + b\sqrt{a}))^{\top}$ as its only attractor, and a saddle $x^{u} = (\sqrt{a},\, 2(1 - b\sqrt{a}))^{\top}$. This has been chosen deliberately simple but is still sufficient to demonstrate all four problems. This way, we do not have to cope with problems inherent to the example system, like high-dimensionality or chaos.

Its flow is shown in figure 1(a). For a chosen trajectory starting at $x^{a} = (2.8,\, 6.2)^{\top}$ (see figure 1(a)) the time-dependence of the Euclidean distance to the attractor

$$d_{\mathrm{E}}(\varphi(t,\, x^{a}),\, x^{\mathrm{s}}) = \sqrt{\varphi(t,\, x^{a})^{\top} \cdot x^{\mathrm{s}}} \tag{4}$$

is depicted in figure 1(a). Two common metrics are the times when an ϵ-neighborhood is entered the first and the last time. So we define the class of sets

$$T^{\epsilon}(x^{a}) = \{t \mid \epsilon = d_{\mathrm{E}}(\phi(t,\, x^{a}),\, x^{\mathrm{s}})\}, \tag{5}$$

that invert the axes of figure 1(b) as depicted in figure 1(c) (blue dotted line). $T^{\epsilon}(x^{a})$ is the set of times when the ϵ-neighborhood is entered or left. Furthermore, the first and last entry times are then $T_{\mathrm{F}}^{\epsilon}(x^{a}) = \inf T^{\epsilon}(x^{a})$ and $T_{\mathrm{L}}^{\epsilon}(x^{a}) = \sup T^{\epsilon}(x^{a})$ respectively. They are graphed in figure 1(c) also.

The infinite reaching time (problem (I)) is visible in figure 1(c) right away, as $T_{\mathrm{F}}^{\epsilon}(x^{a}),\ T_{\mathrm{L}}^{\epsilon}(x^{a}) \longrightarrow \infty$ for $\epsilon \longrightarrow 0$. By definition, this implies that all elements in $T^{\epsilon}(x^{a})$ will approach ∞ also.

Problem (II): $T_{\mathrm{F}}^{\epsilon}(x^{a})$ and $T_{\mathrm{L}}^{\epsilon}(x^{a})$ depend heavily on the choice of ϵ. So a proper physical interpretation is rather difficult. The notions of 'close to the attractor' or 'when the transient is over' depend strongly on ϵ.

The strong discontinuities (problem (III)) for $T_{\mathrm{F}}^{\epsilon}(x^{a})$ and $T_{\mathrm{L}}^{\epsilon}(x^{a})$ when changing ϵ in figure 1(c) make the choice of a proper ϵ even harder. The discontinuities arise because the trajectory will (for a fixed ϵ) enter and exit the corresponding ϵ-neighborhood several times. This behavior is caused by the complex eigenvalues of the system. It could be circumvented locally by choosing a different distance function, for instance

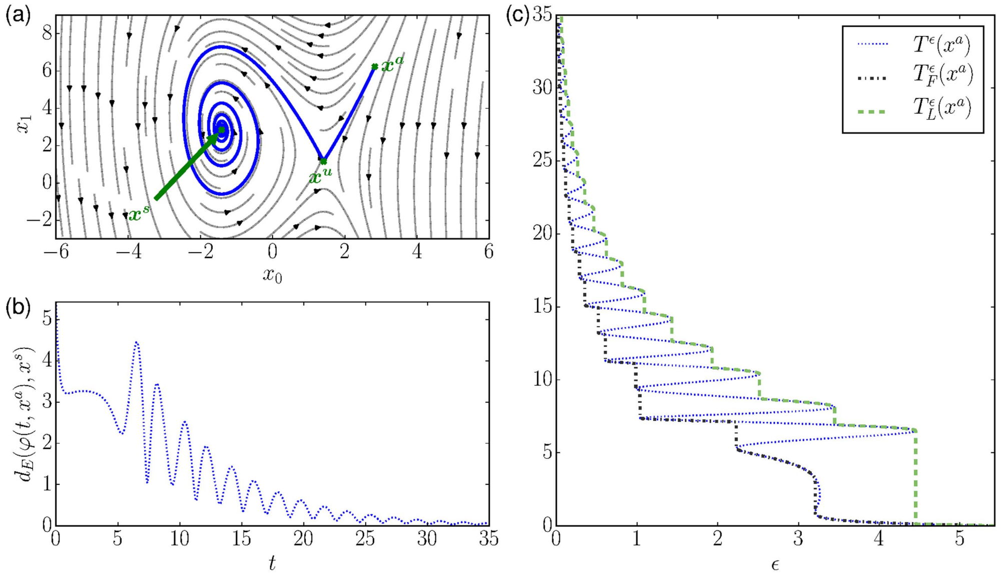


Figure 1. The phase space for the example system equation (3) is depicted in (a). Furthermore, the stable spiraling node x^s and the saddle x^u are added. The trajectory (blue) starting at x^a closely passes by x^u before it finally circulates in to x^s. (b) shows the Euclidean distance d_E (dotted, blue) of this trajectory to its attractor x^s over time t. The first longer dip between $t = 1$ and $t = 5$ is the transient at the (unstable) saddle x^u while the oscillations afterwards are the spiraling around x^s. (c) turns (b) around in order to show the dependence of the time t on some distance $d = \epsilon$ (dotted, blue) of the trajectory to the attractor x^s. Secondly, there are multiple values of ϵ for each t so observables like $t^\epsilon_F(x^a)$ and $t^\epsilon_L(x^a)$ need to be introduced. $t^\epsilon_F(x^a)$ (dash–dotted, black) marks when the first time the ϵ-neighborhood around x^s is entered and $t^\epsilon_L(x^a)$ (dashed, green) the last time. The implications, particularly the arising problems for time definitions, are described in text.

$$d_P(x, x^s) = \|P^{-1} \cdot (x - x^s)\|, \quad P = \begin{pmatrix} 1 & 1 \\ 4\sqrt{a}/\lambda_+ & 4\sqrt{a}/\lambda_- \end{pmatrix}, \tag{6}$$

where $\lambda_\pm = -b/2 \pm \sqrt{b^2/4 - 2\sqrt{a}}$ are the complex eigenvalues of the linearization of (3) around x^s. (This is related to the $\|\cdot\|_P$-norm used in Suppl. Mat. Proposition 3.4.) Unfortunately, this is not so easy for more complex attractors, e.g. the later treated chaotic Rössler system. However, we will present a pragmatic solution to this problem in section 3.2.

Finally, using a different set of coordinates, i.e. smoothly transforming the system, gives different values for $T^\epsilon_F(x^a)$ and $T^\epsilon_L(x^a)$, because the Euclidean distance is not invariant. Hence, the result depends on the set of coordinates chosen for the system and is not invariant under coordinate transformations (problem (IV)). This dependence on the chosen distance is also known to appear in finite-time dynamical systems and their stability [26].

3. Two complementary metrics

To treat the aforementioned problems, we devise two metrics for a general system as equation (1): *area under distance curve* (abbreviated as D) and *regularized reaching time* (T_{RR}). They naturally lead to a transient analysis from separate points of view as explained in the following.

3.1. Area under distance curve

Area under distance curve (D) comes from the idea that a trajectory stays distant from the attractor during the transient while it is close in the asymptotics. A distance function $d(\cdot, \cdot)$ is needed to have notions of 'far' and 'close' and we define D

$$D(x) = \int_0^\infty \mathrm{d}t \; d(\varphi(t, x), \mathcal{A}), \tag{7}$$

where $\mathcal{A}$ is the attractor with the basin $\mathcal{B}_\mathcal{A}$ and φ the time-evolution operator as in equation (2). So we look at the cumulative distance to the attractor and remove the influence of the asymptotics. As (7) is the integral over the distance in time, D is the area below the distance curve. A different point of view is that it is the time weighted by the distance.

As D is defined with the limit of an integral, a note on the convergence is due and can be found in the discussion in section 5.

The distance function d should be between a point in state space and the attractor $\mathcal{A}$. If $\mathcal{A}$ contains more than just one element, it could be the infimum of the distances to all points within. Choosing a tailor-made function $d(\cdot, \cdot)$ allows to adapt the metric to specific research questions, e.g. by letting $d(x, \mathcal{A})$ represent some form of costs or damages due to being away from the attractor. For the idea of D to work, d needs to approach 0 around the attractor and be 0 on it.

Due to the integral representation, D can be estimated numerically directly from the trajectory assuming the attractor is known. The latter was taken as a prerequisite for this article as we want to emphasize the analysis of the transient.

Initial conditions with relatively high values of D are called 'reluctant' and those with low values 'eager'. This terminology is used to emphasize that reluctant states go through large transients distant from the attractor, while eager states approach it directly.

By straightforward differentiation, we can compute the orbital derivative

$$\frac{\partial}{\partial t} D(\varphi(t, x)) = -d(\varphi(t, x), \mathcal{A}), \tag{8}$$

meaning its value strictly decreases along the flow; a property we use later. Furthermore, this shows it to be a Lyapunov function [27]. Furthermore, using equation (8) and adding the condition $D(x) = 0 \;\; \forall x \in \mathcal{A}$ is an alternative definition for D.

3.2. Regularized reaching time

The second idea, *regularized reaching time* (T_{RR}), is based on time differences between trajectories. It can be interpreted as the additional time (positive or negative) that a trajectory starting at a point of interest needs to approach the attractor after a reference trajectory has already approached it. A positive or negative value means that the trajectory at hand approaches the attractor by this much later or earlier, respectively, than the reference trajectory does.

To formalize this idea, we introduce $t^{\epsilon}(x)$ as the time a trajectory starting at an arbitrary state x needs in order to reach an ϵ-environment around the attractor. That means for some function $\Delta : X \longrightarrow \mathbb{R}_{\geqslant 0}$ it holds that

$$\epsilon = \Delta(\varphi(t^{\epsilon}(x), x)), \tag{9}$$

where we want Δ to be 0 on the attractor and $\Delta(\varphi(t, x))$ to be strictly and continuously decreasing in t. This means, in equation (9), Δ has the role of a generalized distance function, measuring how far a point in state space is away from the attractor. Note that $\varphi(t^{\epsilon}(x), x)$ is the state after starting at x and evolving the system for a time $t^{\epsilon}(x)$. Hence, equation (9) implicitly defines $t^{\epsilon}(x)$ to be the time at which an ϵ-environment around the attractor, with respect to the generalized distance function Δ, is entered.

Since the actual reaching times to the attractor are both infinite, T_{RR} is formally described as the limit for $\epsilon \longrightarrow 0$ of the difference between how long the trajectory starting at some arbitrary state x and the trajectory starting at a chosen (fixed) reference point x^{ref} need to enter the corresponding ϵ-environment

$$T_{\mathrm{RR}}(x; x^{\mathrm{ref}}) = \lim_{\epsilon \to 0} (t^{\epsilon}(x) - t^{\epsilon}(x^{\mathrm{ref}})). \tag{10}$$

For hyperbolic fixed points, we prove in Suppl. Mat. section 3 under mild conditions that there exists a class of choices for Δ such that this limit exists for all x within the basin of attraction except the strong stable manifold and the attractor itself. We call the manifold associated to all Lyapunov exponents except the leading one the strong stable manifold. And, if T_{RR} exists, it is unique, i.e. independent of which Δ has been chosen from the class.

Furthermore, we show that T_{RR} is a parametrization of the strong stable foliation. Thus, after a smooth change of coordinates Φ, i.e. a diffeomorphism of the state space, the diffeomorphic image of the strong stable foliation will again parametrize the level sets of T_{RR} in the new variables. Therefore, T_{RR} is invariant under such transformations and it holds that

$$T_{\mathrm{RR}}(\Phi(x); \Phi(x^{\mathrm{ref}})) = T_{\mathrm{RR}}(x; x^{\mathrm{ref}}), \tag{11}$$

where we obviously transformed x^{ref}, too.

T_{RR} represents *the actual time* by how much a trajectory approaches the attractor later or earlier than the one starting at the reference point, so we call states with relatively low T_{RR} 'early' and with high T_{RR} 'late'.

Different choices of x^{ref} (that are not on the strong stable manifold or the attractor) result in additive constants. To be precise, choosing another $x^{\mathrm{ref}\prime}$ yields

$$T_{\mathrm{RR}}(x;\,x^{\mathrm{ref}}) - T_{\mathrm{RR}}(x;\,x^{\mathrm{ref}\prime}) = T_{\mathrm{RR}}(x^{\mathrm{ref}\prime};\,x^{\mathrm{ref}}). \tag{12}$$

Because the rhs of equation (12) does not depend on x, different choices of x^{ref} do not influence the structure of T_{RR} w.r.t. x. Thus central moments, i.e. ones invariant under shifts, are sensible for analyzing T_{RR} over a distribution of initial conditions in state space; especially the standard deviation proves useful for the examples below. In particular, for any choice of x^{ref} it obviously holds that $T_{\mathrm{RR}}(x^{\mathrm{ref}};\,x^{\mathrm{ref}}) = 0$.

The reference point should not be chosen on the attractor because this gives $t^{\epsilon}(x^{\mathrm{ref}}) = 0$ for any ϵ, but for $x \in X\backslash\mathcal{A}$ the time $t^{\epsilon}(x) \to \infty$ for $\epsilon \to 0$. Vice versa, this means when having chosen $x^{\mathrm{ref}} \not\subseteq \mathcal{A}$ then $T_{\mathrm{RR}}(x;\,x^{\mathrm{ref}}) = -\infty\ \ \forall x \in \mathcal{A}$. The same holds for the strong stable manifold.

In order to compute the orbital derivative $\frac{\partial}{\partial t}T_{\mathrm{RR}}(\varphi(t,\,y);\,x^{\mathrm{ref}})$, we use equation (10) and find

$$\frac{\partial}{\partial t}T_{\mathrm{RR}}(\varphi(t,\,y);\,x^{\mathrm{ref}}) = \lim_{\epsilon\to 0}\frac{\partial}{\partial t}t^{\epsilon}(\varphi(t,\,y)), \tag{13}$$

where $y \in X$ is an arbitrary state and exchangeability of the limit and the derivative has been assumed. Next, we take the derivate with respect to time t in equation (9) for $x = \varphi(t,\,y)$. Sorting the terms appropriately gives

$$0 = \left(\frac{\partial}{\partial t}(\Delta\circ\varphi)\right)(t^{\epsilon}(\varphi(t,\,y)) + t,\,y)\cdot\left(\frac{\partial}{\partial t}t^{\epsilon}(\varphi(t,\,y)) + 1\right). \tag{14}$$

$\Delta(\varphi(t,\,x))$ is strictly decreasing in t for any $x \in X$. So its derivative is $\frac{\partial}{\partial t}\Delta(\varphi(t,\,x)) = \left(\frac{\partial}{\partial t}(\Delta\circ\varphi)\right)(t,\,x) < 0$ and in particular non-zero. Hence, $\frac{\partial}{\partial t}t^{\epsilon}(\varphi(t,\,y)) = -1$, leading finally to the orbital derivative of T_{RR}

$$\frac{\partial}{\partial t}T_{\mathrm{RR}}(\varphi(t,\,y);\,x^{\mathrm{ref}}) = -1. \tag{15}$$

This equation is actually rather natural, as the change of time to approach the attractor along the trajectory should exactly be the time passed. Also, this makes it a Lyapunov function [27].

To use equation (15) as an alternative definition we need another constraint. Because of $T_{\mathrm{RR}}(x;\,x^{\mathrm{ref}}) = -\infty\ \ \forall x \in \mathcal{A}$, this cannot be done on the attractor (in contrast to D). In case of hyperbolic fixed points, it follows directly from Suppl. Mat. Proposition 3.7 that T_{RR} is a parametrization of the strong stable foliation $\mathcal{F}_{\mathrm{ss}}$, whose definition is recalled in Suppl. Mat. Theorem 3.6. So we can use the constraint that $T_{\mathrm{RR}}(x;\,x^{\mathrm{ref}}) = 0\ \ \forall x \in \mathcal{F}_{\mathrm{ss}}^{\mathrm{ref}}$, where we call $\mathcal{F}_{\mathrm{ss}}^{\mathrm{ref}} = \mathcal{F}^{\mathrm{ss}}(x^{\mathrm{ref}})$ the *reference leaf* containing x^{ref}. For more complex attractors, a generalized condition needs to be found and this is part of the outlook.

Suppl. Mat. Proposition 3.4 provides the convergence of T_{RR} in equation (10) for hyperbolic fixed points only. When thinking about more complex attractors that may arise in real-world examples the question of convergence comes up again. A general idea why T_{RR} should converge with a well chosen Δ in this case, too, is that in the asymptotics, trajectories will 'behave similarly' because they are close to the attractor. So, for two very small $\epsilon_1 > \epsilon_2$, the time difference to enter the ϵ_2-environment after entering the one of ϵ_1 should be roughly the same, independent from where a trajectory started. Hence, for two states x and x^{ref} we can assume $t^{\epsilon_2}(x) - t^{\epsilon_1}(x) \approx t^{\epsilon_2}(x^{\mathrm{ref}}) - t^{\epsilon_1}(x^{\mathrm{ref}})$ implying $t^{\epsilon_2}(x) - t^{\epsilon_2}(x^{\mathrm{ref}}) \approx t^{\epsilon_1}(x) - t^{\epsilon_1}(x^{\mathrm{ref}})$. This suggests that the limit in equation (10) might exist. So a crucial problem is to find an appropriate function for Δ in order to get an estimation for T_{RR}.

Estimation of T_{RR}. The first idea for a Δ would be the infimum of the Euclidean distance to the points in the attractor. Basically, this means that t^{ϵ} should be replaced by $T_{\mathrm{F}}^{\epsilon}$ or $T_{\mathrm{L}}^{\epsilon}$ from section 2. This would give a very coarse estimation but is probably not the correct choice as the condition of Δ being strictly decreasing along the flow is in general not fulfilled.

A pragmatic choice of Δ is D, the area under distance curve. It fulfills both conditions demanded for Δ (see section 3.1) when using for d the infimum of the Euclidean distance to the attractor points. Hence, we can define $t_D^{\epsilon}(x)$ as the time until the D (equation (7)) of the trajectory's remainder is ϵ-small

$$\epsilon = D(\varphi(t_D^{\epsilon}(x),\,x)) = \int_{t_D^{\epsilon}(x)}^{\infty}\mathrm{d}t\ d(\varphi(t,\,x),\,\mathcal{A}). \tag{16}$$

Note that the ideas for D and T_{RR} are generally independent and the usage of D in this case is purely because it fulfills the above mentioned conditions. So it is a good, pragmatic choice.

Using t_D^{ϵ} defined in equation (16) as the time-function t^{ϵ} in equation (10) for the estimation of T_{RR}, our numerical results show that this idea is sensible for more complex attractors, e.g. in the Rössler system below.

4. Examples

In order to demonstrate the applicability of the metrics, we selected four examples with differing properties and increasing complexity.

4.1. Linear system with two different time scales

Even though we want to focus on going in the direction of application to real-world systems, understanding some features in a basic linear system proves useful. For general systems, T_{RR} and D can be tackled numerically only. But a linear system can be solved analytically and explicit expressions for both metrics were found. We will first analyze both metrics for a general linear system and then discuss a chosen example.

T_{RR} for a general linear system. For a hyperbolically stable linear system with a (complex-)diagonalizable matrix $A \in \mathbb{R}^{n\times n}$ and the fixed point x^{f} at the origin,

$$\dot{x} = A \cdot x, \tag{17}$$

we decompose $x = \sum_{i=0}^{n-1} \alpha^i v^i$ with coefficients $\alpha^0,\ldots,\alpha^{n-1}$ in the eigenvector basis $v^0,\ldots,v^{n-1}$ with eigenvalues $\lambda^0,\ldots,\lambda^{n-1}$ sorted in descending order by real part. We assume in particular λ^0 to have a strictly larger real part than λ^1 and multiplicity one. Hence we can apply Suppl. Mat. equation (10) derived in the Suppl. Mat. and get

$$T_{\mathrm{RR}}(x;\, x^{\mathrm{ref}}) = \frac{1}{\lambda_0} \ln \left| \frac{\alpha^{0,\mathrm{ref}}}{\alpha^0} \right|, \tag{18}$$

where $\alpha^{0,\mathrm{ref}}$ is the α^0 coefficient for the reference point x^{ref}. $\alpha^{0,\mathrm{ref}}$ should be non-zero, i.e. x^{ref} should not be on the strong stable manifold.

Note that Suppl. Mat. Proposition 3.4 gives the uniqueness of this result independent of the choice of Δ.

In equation (18), T_{RR} depends only on α^0, meaning the projection of x on the eigenvector corresponding to the least stable eigenvalue λ^0. While this might be counter-intuitive in the beginning, it can be explained: the contributions from all other eigenvalues are vanishing because they decay faster than λ^0 by definition. So for a linear system, only the contribution from λ^0 remains. Also, on the strong stable manifold where $\alpha^0 = 0$, the values for T_{RR} go to $-\infty$ which we mentioned already in section 3.2 for general systems.

D for a general linear system. Taking the system (17) and choosing $d(x,\, \{x^{\mathrm{f}}\}) = d_E(x,\, x^{\mathrm{f}})^2$ the squared Euclidean distance, we calculate D directly by using the definition equation (7)

$$D(x) = \sum_{i,j=0}^{n-1} \frac{-(\alpha^i)^* \alpha^j}{(\lambda^i)^* + \lambda^j} (v^i)^\dagger v^j. \tag{19}$$

Therefore, in case of D, all eigenvalues contribute, contrary to T_{RR}. But they are weighted as can be seen in the denominator. In case of A being symmetric, this formula can be reduced to $D(x) = \frac{1}{2} x^\top A^{-1} x$.

T_{RR} for an example linear system. We choose the $n =$ two-dimensional linear system

$$\dot{x} = \begin{pmatrix} -1 & 0 \\ 4 & -2 \end{pmatrix} \cdot x \tag{20}$$

with a stable and a strong stable eigenvalue and corresponding eigenvectors

$$\lambda^{\mathrm{s}} = -1,\ v^{\mathrm{s}} = \begin{pmatrix} 1 \\ 4 \end{pmatrix} \text{ and } \lambda^{\mathrm{ss}} = -2,\ v^{\mathrm{ss}} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \tag{21}$$

We choose the reference point to be $x^{\mathrm{ref}} = (1,\, 1)^\top$. Identifying $\lambda^0 = \lambda^{\mathrm{s}}$, $v^0 = v^{\mathrm{s}}$ implies $\alpha^0 = x_0$. Then, using equation (18) gives

$$T_{\mathrm{RR}}(x;\, x^{\mathrm{ref}}) = \ln(|x_0|). \tag{22}$$

This result is also visible in the numerical estimation in figure 3(c); the values of T_{RR} change only in the direction of x_0. The coloring describes the values of the metrics (see the colorbar in the right of the figures) and the green star represents x^{ref}.

In order to get a better feeling for these metrics, we have chosen two exemplary initial conditions, an *early-eager* one and a *late-eager* one, and plotted their trajectories' distance to the attractor over time in figure 2. We see an intuition for T_{RR}: it can be interpreted as the time-shift between the original trajectory and the reference trajectory until the asymptotics match. So we plotted both trajectories shifted to each other using the analytical result for T_{RR} in equation (22).

D for an example linear system. Analyzing D for the example linear system in equation (20) gives

$$D(x) = \frac{11}{6} x_0^2 + \frac{1}{4} x_1^2 + \frac{2}{3} x_0 x_1, \tag{23}$$

where equation (19) has been used. The numerical result in figure 3(a) confirms this.

In figure 2, the blue-shaded area corresponds to the D value which is the same in both cases of our particular choice. This choice was made in order to see how trajectories can have differing T_{RR} values even if the D values match.

The exponential lower bound that comes up in the scatter plot figure 3(b) can be calculated analytically by combining equations (22) and (23)

Reproduced from: T. Kittel et al., Timing of transients: Quantifying reaching times and transient behavior in complex systems, New J. Phys., vol. 19, no. 8, 2017, doi: 10.1088/1367-2630/aa7b61. Published under Creative Commons Attribution License 3.0 (CC BY).

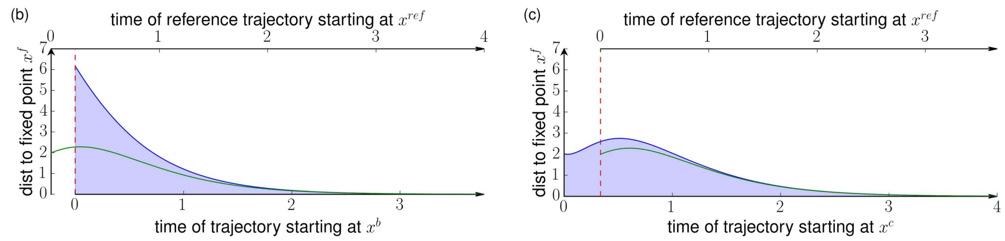


Figure 2. The figure shows for two exemplary initial conditions (a) $x^b = (0.8,\ 2.35)^{\top}$ and (b) $x^c = (1.4,\ 0.24)^{\top}$ the distance of the attractor over time (blue curve) in the linear example system of section 4.1. The initial conditions have been chosen such that the D value, which corresponds to the blue-shaded area, is the same for both trajectories, $D(x^b) = D(x^c) = 3.8$. But the trajectory starting at x^b approaches the attractor *earlier* than the reference trajectory (green in (a) and (b)), which in turn is *earlier* than the one from x^c, meaning $T_{\mathrm{RR}}(x^b;\ x^{\mathrm{ref}}) = -0.22 < T_{\mathrm{RR}}(x^{\mathrm{ref}};\ x^{\mathrm{ref}}) = 0 < T_{\mathrm{RR}}(x^c;\ x^{\mathrm{ref}}) = +0.34$. In order to show this, the example trajectories (blue) have been shifted in each plot by the value of T_{RR} with respect to the reference trajectory (green). This demonstrates an intuition behind T_{RR}: it describes by how much one has to shift one trajectory so it matches the asymptotics of the reference trajectory.

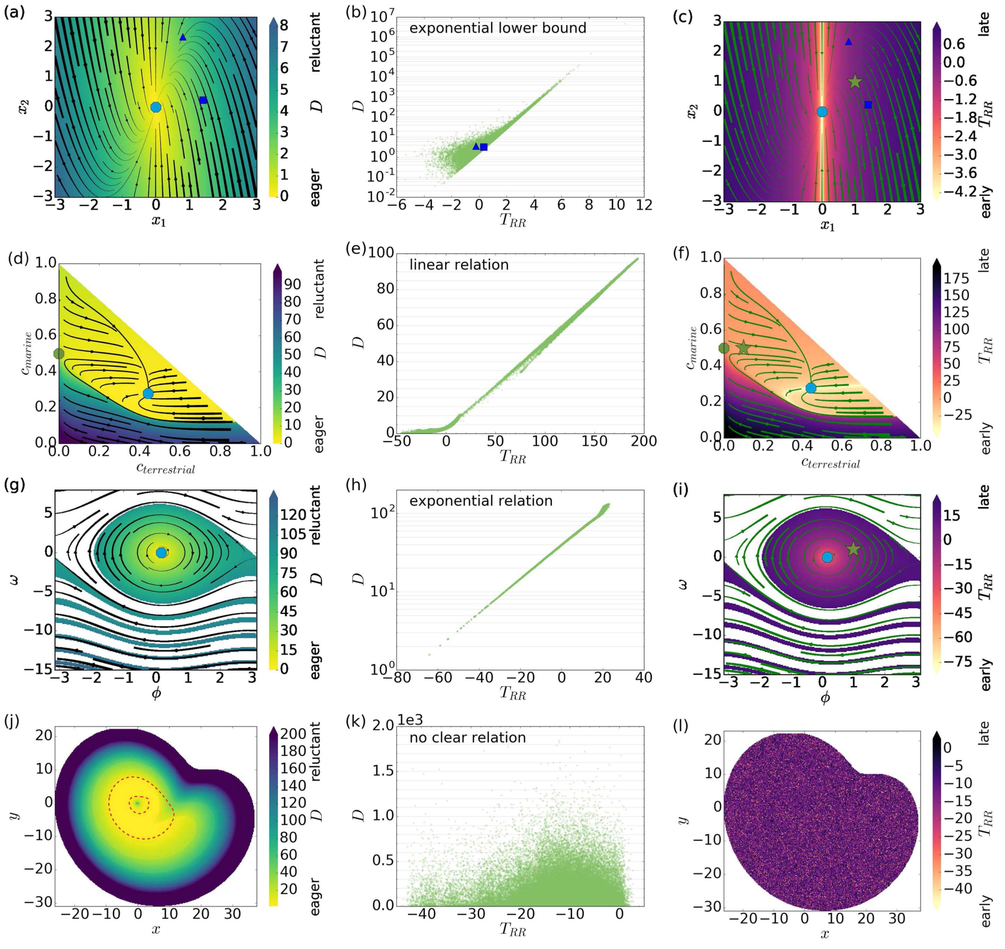


Figure 3. For the presented example systems (top to bottom: linear system, global carbon cycle, generator in a power grid, Rössler system) the two new metrics have been computed for each initial condition in the state space and marked with color, see left column *area under distance curve* (D, D) and right column *regularized reaching time* (T_{RR}). The middle column shows their relations for the particular system. The initial conditions x^b (triangle) and x^c (square) from figure 2 have been marked in (a)–(c), too. As the Rössler system is three-dimensional, the above plot depicts only a slice at fixed $z = 0.6$ where the boundary of the attractor's projection to this plane is shown in dashed red lines. As this is only a projection, D is not 0 for all points within. For comparison, a graphical representation of the full attractor can be found in supplementary material section 2.

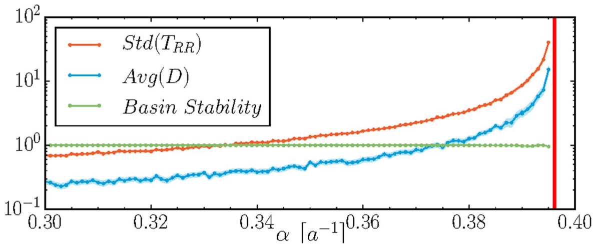


Figure 4. For the *global carbon cycle* in equations (25*a*) and (25*b*), the mean of D and standard deviation of T_{RR} are plotted (with their 5%–95% bootstrap error) and show a divergence before the parameter α (yearly human carbon offtake) reaches the bifurcation value (marked by the red line). For comparison Basin stability is shown, which does *not* show any change because the size of the basin stays constant before the bifurcation.

$$D(x) \geqslant \frac{25}{18}\mathrm{e}^{2T_{\mathrm{RR}}(x;x^{\mathrm{ref}})}. \tag{24}$$

4.2. Global carbon cycle

The second example has been chosen to take a step in the direction of real-world examples. It is a conceptual model of the global carbon cycle proposed by Anderies *et al* [15]. We use the pre-industrialization version. It consists of three dynamical variables, the terrestrial, marine and atmospheric carbon stocks, denoted by $c_{\mathrm{t}} = c_{\mathrm{terrestrial}}$, $c_{\mathrm{m}} = c_{\mathrm{marine}}$ and $c_{\mathrm{a}} = c_{\mathrm{atmospheric}}$ respectively. Furthermore, the conservation of total carbon is formulated in the constraint $C = c_{\mathrm{t}} + c_{\mathrm{m}} + c_{\mathrm{a}} = \mathrm{const}$. Thus, we can reduce the system to 2 state variables c_{t} and c_{m} and rescale the units such that $C = 1$:

$$\dot{c}_{\mathrm{t}} = \mathrm{NEP}(p, r, c_{\mathrm{t}}) - \alpha c_{\mathrm{t}}, \tag{25a}$$

$$\dot{c}_{\mathrm{m}} = I(c_{\mathrm{a}}, c_{\mathrm{m}}), \tag{25b}$$

where NEP is the net Eco-system production, p photosynthesis, r respiration, α harvesting parameter and I diffusion; indirect dependencies have been omitted and more details are in [15, 28]. As the full equations are rather lengthy, we put them in Suppl. Mat. section 1 and refer in the analysis to the flow that is drawn in figures 3(d) and (f) and the α parameter stated above. The whole phase space of equations (25*a*) and (25*b*) is the basin of the attraction of the fixed point in the middle marked by a blue dot; the dynamics is drawn as streams. The trajectories starting in the lower part have to pass by a 'desert-like' saddle (with $c_{\mathrm{t}} = 0$) at the left (green dot).

The color in figure 3(f) depicts T_{RR} and the first finding is the splitting of the basin of attraction. The strong stable manifold of the stable node becomes visible as a light beige line due to its low values of T_{RR}, i.e. as very *early* states because $T_{\mathrm{RR}} \to -\infty$. So it is the separatrix for the observed splitting. Also, the expected smooth increase of the return times when distancing (along the trajectories) from the attractor can be observed.

Still, the splitting of the basin of attraction is visible for values of $c_{\mathrm{terrestrial}} < 0.3$, where it is only due to quantitatively different behavior and the visible boundary is actually a rather sharp but still continuous transition. (The latter statement follows right from Suppl. Mat. Theorem 3.6 and Suppl. Mat. Proposition 3.7.) Looking at figure 3(f) one can also see that the boundary becomes more and more fuzzy for even smaller values of $c_{\mathrm{terrestrial}}$, demonstrating that there is really a need for a quantitative analysis.

When applying D to this model (figure 3(d)), the splitting of the basin can be observed again. In contrast to T_{RR}, the strong stable manifold of the stable node is not visible because D can be seen as a (by distance) weighted time and the contributions from the asymptotic part where the difference in the Lyapunov spectrum matters are negligible.

Furthermore, we see a clear linear correlation of both metrics in figure 3(e) because all trajectories starting in the lower part have to pass by at the saddle on the left and spend a long time there.

Both metrics work as *early-warning signals* [14, 29], too. When increasing α, corresponding to the harvest of terrestrial carbon, the system passes through a subcritical pitchfork bifurcation where the saddle becomes stable and the lower-left part of the phase space splits off. The divergences of the two metrics' statistics as seen in figure 4 prove their prebifurcational sensitivity, while other systemic indicators like basin stability [30] do not change (up to numerical fluctuations, see figure 4). Note that in this example, a Lyapunov exponent analysis of the saddle would be able to predict the bifurcation due to the simplicity of the saddle also. However, in case of a more complex saddle, this would become arbitrarily difficult while this numerical estimation would still be possible for both metrics.

4.3. Generator in a power grid

As the next example, we chose the swing equation in equation (26), a basic model describing the dynamics of a single generator connected to a large power grid [31]. It consists of two dynamical variables, the phase θ and angular frequency ω, both in a reference frame rotating at the grid's rated frequency. The parameters of the system correspond to the net power production $P = 1$ (at the node), the capacity of the transmission line $K = 6$ and dampening $\alpha = 0.1$.

$$\dot{\phi} = \omega, \quad \dot{\omega} = 2P - \alpha\omega - 2K \sin\phi. \tag{26}$$

In this form, which is used in electrical engineering [25, 32], it is formally equivalent to a pendulum with constant driving and damping.

The stable fixed point at $\omega^{\mathrm{s}} = 0, \phi^{\mathrm{s}} = \arcsin\frac{P}{K}$ describes a state of synchronization. For the chosen set of parameters, the system exhibits another attractor: a limit cycle at larger positive values of ω. For negative values, the two basins of attraction are interleaved. A more detailed introduction and analysis can be found in [25, 31, 33].

Calculating T_{RR} inside the basin of the stable fixed pointed $(\omega^{\mathrm{s}}, \theta^{\mathrm{s}})$ yields figure 3(i). There is basically no color change away from the attractor, so we can see that a trajectory barely spends any time in the transient and goes quickly to the attractor. Analogously, figure 3(g) for D leads to the same conclusion as T_{RR}.

Comparing both metrics in figure 3(h) shows that they are closely linked. Note that this time D is presented on a logarithmic scale, so the relation is exponential and what we see here is actually the influence of the linearized part of the system. The accumulation in the upper right corresponds to the initial conditions with lower values of ω. This means, they only go through a very short transient and spend most of their time in the part where the linearization holds.

The white parts in the phase spaces figures 3(g) and (i) correspond to the basin of attraction of the limit cycle. As this means the system is away from synchrony, the generators would usually switch off before reaching it. So we did not include it in the analysis.

4.4. Chaotic Rössler oscillator

Although we have proven the convergence of T_{RR} for fixed points only, we show with the chaotic Rössler system [34, 35] that both metrics are applicable to higher-dimensional and more complex attractors, too. The equations are

$$\dot{x} = -y - z, \quad \dot{y} = x + ay, \quad \dot{z} = b + z(x - c), \tag{27}$$

where x, y and z are the coordinates in state space. While this naming convention is not in line with the rest of the article, it has been chosen as it is standard for these equations.

Figure 3(l) shows a slice of the phase space with the standard parameters $a = 0.2, b = 0.2, c = 5.7$ for T_{RR} and the expected sensitivity to initial conditions for chaos is observed: *early* and *late* trajectories lie closely together and the metric T_{RR} has low spatial correlation.

In contrast, D shows in figure 3(j) surprisingly smooth changes of an embryo-like shape. Because the focus of this article is on transient dynamics a new feature of the chaotic Rössler system is uncovered: while the attractor is chaotic, the basin of attraction is very regular. D focuses on the initial transient and the chaotic asymptotics is filtered out. For comparison, the boundaries of the attractor's projection have been added with dashed red lines in figure 3(j) and depictions of the attractor are in Suppl. Mat. section 2.

Furthermore, T_{RR} can be applied as an early-warning signal in this case, too. In order to demonstrate this, we chose to vary a as it has a crucial influence on the system's dynamics (see the bifurcation diagram in figure 5 (green)). For values of $a < 0.006$ (see [36]) there is only a single stable fixed point. At $a \approx 0.006$ a limit cycle emerges due to a Hopf bifurcation [36]. For $a > 0.11$, several period doublings are observed, finally leading to chaos for $a > 0.155$. Even in the chaotic regime, further bifurcations can be observed.

In figure 5, the standard deviation of the T_{RR} distribution from randomly chosen initial conditions inside the basin of attraction is given. Due to the sensitive dependence on initial conditions, the reference value varies a lot and hence introduce shifts in the distribution that do not describe actual changes in the system's dynamics. To remove this effect, it is crucial to use central moments like the standard deviation.

T_{RR} is strongly sensitive to any qualitative changes in the dynamics of the system, incl. even chaos–chaos transitions. Closely observing figure 5 uncovers that there is a base-line with little fluctuations at $\mathrm{Std}(T_{\mathrm{RR}}) \approx 10$ complemented with strong peaks. In the chaotic regime, the peaks correspond directly to qualitative changes. Also, we observe sensible changes during the period-doubling phase and a strong increase before the Hopf bifurcation at $a \approx 0.006$, proving the usefulness as an *early-warning signal*.

The abrupt downward peak at $a \approx 0.11$ is unexpected and more details are needed to clarify it. The other peaks correspond well with the transitions visible in the bifurcation diagram.

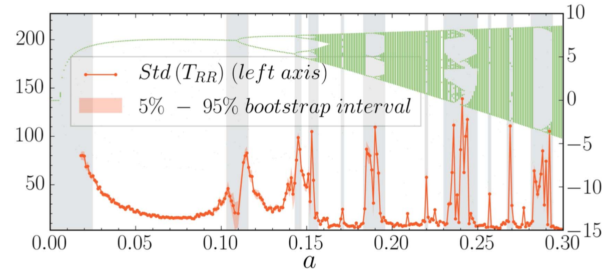


Figure 5. The bifurcation diagram (green) of the Rössler system for varying the parameter a in equation (27) was computed from the local maxima in z of the attractor and T_{RR} (orange) shows a strong sensitivity to these qualitative changes. The gray background is used so the reader can more easily connect the peaks in T_{RR} to the corresponding parts in the bifurcation diagram.

5. Discussion

In order to see how far the two proposed metrics answer the question 'How can we properly quantify the time to reach a system's attractor?' we will go along the four essential problems that have been worked out in section 2 for this discussion: (I) infinite reaching time, (II) physical interpretation (III) discontinuities and (IV) non-invariance.

Area under distance curve (D) has been defined as the cumulative distance to the attractor over time in order to emphasize the idea that a trajectory stays 'far' from the attractor in the transient while being close in the asymptotics. The distance d is not necessarily meant in the mathematical sense [37], but it only needs to approach 0 around the attractor and be 0 on it. In that way, it is possible to choose the appropriate d for different research questions, e.g. asking about costs or damages. Even in these interpretations D is a metric capturing the transient time, because there are only contributions when the trajectory is distant from the attractor, i.e. still in its transient phase. Another point of view is to see D as the time to reach the attractor weighted by the distance.

We understand Problem (I), infinite reaching time, as solved. For hyperbolic attractors and d being a mathematical distance function, the integral in equation (7) does converge. Trajectories approach the attractor exponentially in the asymptotics and the integral over the exponential envelope is finite.

While this covers most systems relevant for real-world applications, in some very specific cases, D might be infinite. The asymptotic tail of the integral might not converge, i.e. the trajectory does not approach it 'fast enough'. This means, either this is the wanted result or d has not been chosen appropriately. In the first case, it could be for example that D was computed for an initial condition that is not economically feasible, so the cost diverges. Furthermore, this would imply that even though the attractor is systemically stable, it is not economically feasible to cope with small perturbations.

From a technical perspective a divergence in D can be understood as indicating that d has not been chosen matching to the system. E.g. using the Euclidean distance and $\dot{x} = -\frac{1}{2}x^3$ where the solutions are $\varphi(t,\, x) = \mathrm{sign}(x)\frac{1}{\sqrt{t + |x|^{-2}}}$, D does not converge. Another example is to take a linear system $\dot{x} = -x$ with $x < 1$. Using $d(x,\, \{0\}) = -\frac{1}{\ln|x|}$ with $d(0,\, \{0\}) = 0$ gives $D \to \infty$.

This can usually be solved by choosing an appropriate d. E.g. choosing for $\dot{x} = -\frac{1}{2}x^3$ using $d(x,\, \{0\}) = \exp(-|x|^{-1})$ and $d(0,\, \{0\}) = 0$ gives finite values for D.

Problem (II) is solved because there is no direct parameter. Still, as there is the indirect dependence on d a discussion is necessary and given in comparison to the first and last entry time to an ϵ-environment $T_{\mathrm{F}}^{\epsilon}(x)$ and $T_{\mathrm{L}}^{\epsilon}(x)$ respectively. For them, a small change in ϵ will have a huge impact on the measured times because for $\epsilon \to 0$ both values go to infinity. Furthermore, if one would locally change the way how the distance to the attractor is measured, the values for $T_{\mathrm{F}}^{\epsilon}(x)$ and $T_{\mathrm{L}}^{\epsilon}(x)$ would change drastically, too.

Because D is defined as the cumulative d over time, a local change in d will have only minor effects on the exact value, so even estimated functions for d with some uncertainty can be used.

Problem (III), discontinuities, have been avoided in D by using the integral representation. Hence the function is even differentiable along the flow (see equation (8)).

We see Problem (IV), non-invariance, as solved, if d has been chosen with some meaning, e.g. economic damages. Then one can simply represent the economic damage function in the changed coordinates, because the meaning is independent of the coordinates. This reasoning is not mathematical but context-dependent. From a purely mathematical point of view, if d is just any distance function, generally the result is not invariant under change of coordinates as it depends on geometric features of the system. But as we want to go in the direction of real-world systems, a model-specific choice of d is compulsory anyway.

Regularized reaching time (T_{RR}) has been defined as the difference in time to approach the attractor.

Problem (I), infinite reaching time, does not appear for all states in the basin of attraction except the attractor and the strong stable manifold. In case of the attractor, the trajectory will stay on it while trajectories from other points approach the attractor only; and, by definition, points on the strong stable manifold approach the attractor a lot faster, also in the asymptotics. So these infinities are actually reasonable results. Also, both are usually of a smaller dimension than the state space. Hence coincidently being there is unlikely, and these cases are rather irrelevant for real-world applications.

Problems (II) and (III) are intrinsically solved by avoiding parameters. The necessary choice of x^{ref} introduces a constant shift only while not changing the structure of the function. When looking at central moments of T_{RR}, i.e. ones invariant under shifts, this dependence on the choice of x^{ref} disappears completely as they would only shift the mean. So an analysis by changing the system's parameters is possible. This has been done in the examples for the global carbon cycle and Rössler system and T_{RR} has been confirmed as an early-warning signal. This analysis can be seen as a systemic approach to the concept of *critical slowing down* (CSD) [14, 29, 38] after a shock, i.e. an instantaneous and non-infinitesimal perturbation, uncovering prebifurcational changes in the transient behavior. In contrast, CSD is usually done with (local) noise only. The usage of shocks has been developed in the context of Basin stability [30, 31] and its extensions [23, 39–42].

Problem (IV), non-invariance, is proven to be solved for hyperbolic fixed points. In case of more complex attractors, we can currently only define an estimation of T_{RR} which depends on geometric properties. So invariance might not be given and more research due in that direction. An important step in that direction has been done by writing down the properties of Δ which imply that the necessary way of measuring how a trajectory approaches might not be local (except fixed-points). The used pragmatic choice of $\Delta = D$ demonstrates this as it basically says that the remainder of the trajectory should have an ϵ small value of D only.

An assumption that has been made during the proof of invariance of T_{RR} for hyperbolic fixed points is: the eigenvalue of the RHS's Jacobian with the largest real-part is either unique and with multiplicity 1 or there are two that are complex conjugated to each other. However, this condition is not really constraining because we assume most real world systems fulfill it.

Comparison. The metrics have been applied to several examples and we will discuss a comparison between both metrics here. They are depicted in figures 3(b), (e), (h) and (k) and show different relations, as stated in the figures. The exponential lower bound and the exponential relation for the linear system and swing equation respectively come mostly from the asymptotic behavior, in particular as the linear system does (by definition) not have any nonlinearities. Still the relations are different as the asymptotic behavior differs slightly, too, one being a node the other a focus. This shows that even though we clearly focus on the transient, it is actually important to be aware of the asymptotic behavior, too. And one cannot analyze the former without knowing about the latter.

In contrast, the linear relation for the global carbon cycle really points to the transient behavior only. It is due to the states passing by the 'desert-like' saddle. Finally, there seems to be no clear relation between both metrics for the Rössler example, pointing to the chaotic behavior. Still, both metrics have separately been useful, D demonstrating the smoothness of the basin of attraction and the standard deviation of T_{RR} being sensitive to qualitative changes of the system.

Other methods. When developing this research on measuring times to approach the attractor, we had the impression that there are two more common ideas, additionally to the first and last entry time. We do not intend to have a complete overview of all methods but would like to discuss these two shortly here. This part refers to a general system in the sense of (1).

The first idea is to develop metrics based on characteristic times. These are usually defined as the time until a quantity is reduced to $1/\mathrm{e}$ of its original value [43]. This quantity could be a distance to the attractor or a coordinate. From this definition it already follows that they are subject to problem (IV). Also, even if the quantity is at $1/\mathrm{e}$ of its initial value, the trajectory might still be far away from the attractor and in its transient dynamics. Lastly, taking a one-dimensional linear system and assuming the quantity is the coordinate, the characteristic time is constant for all initial conditions. This is counter-intuitive when thinking about a time to approach the attractor.

The second idea for general systems is to use Lyapunov exponents [44]. They have units of inverse time and are invariant under changes of coordinates. However, they are actually a property of the attractor. So they do not capture the transient but only the asymptotics closely around and at the attractor.

6. Summary and outlook

In this article, we have treated the question: 'How can we properly quantify the time to reach a system's attractor?'

First, we have worked out the four essential problems of quantifying the timing of transients in order to develop two new metrics, *area under distance curve* D and *regularized reaching time* T_{RR}. As the focus of this work is meant to be on making a first step to real-world systems, we have applied the metrics numerically to four chosen examples systems, observing different features. Finally, we have discussed in detail how far the metrics treat the four essential problems.

With this approach, interesting features of the examples have been uncovered. Using the global carbon cycle, we have demonstrated the importance of the transient analysis, as the desert state is only a saddle but nevertheless passing by there would lead to an extinction of humanity. The splitting of the basin of attraction is partially due to the strong stable manifold of the attractor but it continues for lower values of $c_{\mathrm{terrestrial}}$ where it is only due to quantitatively different behavior demonstrating the need for quantitative methods. Particularly interesting is how the (central) statistics of our metrics are a systemic approach to the concept of CSD leading to an interpretation as early-warning signals, which we have demonstrated also. The independence of the choice of reference points has been achieved by the usage of central moments. In case of the generator in a power grid, most of the relevant dynamics seems to be dominated by the linearization of the equations around the focus.

In order to prove the applicability to more complex dynamics, we have used our metrics on the Rössler system, too, and found the smoothness of the attractor's basin with D. As the attractor itself is chaotic, this smoothness is surprising. T_{RR} reacts strongly to the sensitivity to initial conditions of the chaotic system and one might want to ask whether there is a relation to winding numbers when approaching the attractor. Still, its worth is displayed when varying the a parameter. This parameter has strong influence on the Rössler system's dynamics and T_{RR} reacts strongly to the different bifurcations and even the chaos–chaos transitions, proving again its worth as *early-warning-signal*.

We have not performed any comparative analysis with the mentioned first- and last-entry-time approaches because these behave inconsistently and their quantitative results are arbitrary, as discussed at length in section 2.

The detailed discussion on the two metrics have showed that, while they do treat the four essential problems, they do not fully solve them and further investigation is needed. Also, they come from two very different basic ideas so the comparison showed that they really measure independent features but can improve the understanding of a system by combining them. For both metrics, we have showed that they are Lyapunov-functions. While some properties have already been used in the article, these definitions in terms of orbital derivatives may be a rich groundwork for the next steps.

Four directions of immediate future research are due:

(1) Working on the definition of T_{RR} using the Lyapunov function properties. This step is crucial in order to further the understanding of transient analysis and needs to take the attractor into account as well. Hence, the analysis of more complex attractors and basin shapes, e.g. riddled basins, is part of this.

(2) Applying the current definition of the metrics, in particular using the estimation of T_{RR} with $\Delta = D$, to understand the implications and the precise use cases better. Furthermore, their relations to topological structures, e.g. in complex networks [45], need to be worked out in detail. This part, even though complementary, should be done in accordance with the results in (1).

(3) On the numerical side, it is important to introduce more sophisticated methods of Lyapunov function estimations, where a starting point is the work by Giesl and Hafstein [46]. The curse of dimensionality is going to be a problem for network systems, hence methods for estimation of these metrics' statistics in such kinds of systems induce a need for developing specific algorithms.

(4) Comparison of the timing of transients in model output and observation data as the new observable time is now available.

Acknowledgments

The authors thank the anonymous referees for their detailed and constructive feedback.

This paper was developed within the scope of the IRTG 1740/TRP 2011/50151-0, funded by the DFG/FAPESP. The authors acknowledge the support of the European Commission's Marie Curie Fellowship MALCOD (Grant No. 660616). This work was conducted in the framework of PIKs flagship project on coevolutionary pathways (copan). The authors thank CoNDyNet (FKZ 03SF0472A) for their cooperation. The authors gratefully acknowledge the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the Land Brandenburg for supporting this project by providing resources on the high performance computer system at the Potsdam Institute for Climate Impact Research. The authors thank the developers of the used software: Python [47], Numerical Python [48] and Scientific Python [49].

The authors thank Sabine Auer, Karsten Bölts, Catrin Ciemer, Jonathan Donges, Reik Donner, Jasper Franke, Frank Hellmann, Jakob Kolb, Chiranjit Mitra, Finn Müller-Hansen, Jan Nitzbon, Anton Plietzsch Stefan Ruschel, Tiago Pereira da Silva, Francisco A Rodrigues, Paul Schultz, and Lyubov Tupikina for helpful discussions and comments.

References

[1] Barkema G, Marko J and De Boer J 1994 *Europhys. Lett.* **26** 653

[2] Gaulin B D and Spooner S 1987 *Phys. Rev. Lett.* **58** 668–71

[3] Chou Y C and Goldburg W I 1979 *Phys. Rev.* A **20** 2105

[4] Fiutak J and Mizerski J 1980 *Z. Phys.* B **39** 347–52

[5] Tang C, Telle J and Ghizoni C 1975 *Appl. Phys. Lett.* **26** 534–7

[6] Krapivsky P L, Redner S and Ben-Naim E 2010 *A Kinetic View of Statistical Physics* (Cambridge: Cambridge University Press)

[7] Castellano C, Fortunato S and Loreto V 2009 *Rev. Mod. Phys.* **81** 591

[8] Chowdhury D, Santen L and Schadschneider A 2000 *Phys. Rep.* **329** 199–329

[9] Hastings A 2004 *Trends. Ecol. Evol.* **19** 39–45

[10] Van Geest G, Coops H, Scheffer M and van Nes E 2007 *Ecosystems* **10** 37–47

[11] Schaffer W M, Kendall B, Tidd C W and Olsen L F 1993 *Math. Med. Biol.* **10** 227–47

[12] Fisher R S, Boas W v E, Blume W, Elger C, Genton P, Lee P and Engel J 2005 *Epilepsia* **46** 470–2

[13] Fisher F M 1989 *Disequilibrium Foundations of Equilibrium Economics* (Cambridge: Cambridge University Press)

[14] Lenton T M 2011 *Nat. Clim. Change* **1** 201–9

[15] Anderies J M, Carpenter S R, Steffen W and Rockström J 2013 *Environ. Res. Lett.* **8** 044048

[16] Heitzig J, Kittel T, Donges J F and Molkenthin N 2016 *Earth Syst. Dyn.* **7** 21–50

[17] Crutzen P J 2002 *Nature* **415** 23

[18] Steffen W *et al* 2011 *Ambio* **40** 739–61

[19] Waters C N *et al* 2016 *Science* **351** aad2622

[20] Steffen W, Broadgate W, Deutsch L, Gaffney O and Ludwig C 2015 *Anthropocene Rev.* **2** 81–98

[21] Rockström J *et al* 2009 *Ecol. Soc.* **14** 32

[22] Steffen W *et al* 2015 *Science* **347** 1259855–1–10

[23] van Kan A, Jegminat J, Donges J F and Kurths J 2016 *Phys. Rev.* E **93** 042205–1–7

[24] Kuznetsov Y A 2013 *Elements of Applied Bifurcation Theory* vol 112 (New York: Springer)

[25] Weckesser T, Jóhannsson H and Østergaard J 2013 Impact of model detail of synchronous machines on real-time transient stability assessment *Bulk Power System Dynamics and Control-IX Optimization, Security and Control of the Emerging Power Grid, (IREP) Symp.* (IEEE) pp 1–9

[26] Bhat S P and Bernstein D S 2000 *SIAM J. Control Optim.* **38** 751–66

[27] Giesl P 2007 *Construction of Global Lyapunov Functions using Radial Basis Functions* (Berlin: Springer)

[28] Heck V, Donges J F and Lucht W 2016 *Earth Syst. Dyn.* **7** 783–96

[29] Scheffer M, Bascompte J, Brock W A, Brovkin V, Carpenter S R, Dakos V, Held H, Van Nes E H, Rietkerk M and Sugihara G 2009 *Nature* **461** 53–9

[30] Menck P J, Heitzig J, Marwan N and Kurths J 2013 *Nat. Phys.* **9** 89–92

[31] Menck P J, Heitzig J, Kurths J and Joachim S H 2014 *Nat. Commun.* **5** 1–8

[32] Yuan Y, Kubokawa J and Sasaki H 2003 *IEEE Trans. Power Syst.* **18** 1094–102

[33] Schultz P, Heitzig J and Kurths J 2014 *New J. Phys.* **16** 125001

[34] Rössler O E 1976 *Phys. Lett.* A **57** 397–8

[35] Zgliczynski P 1997 *Nonlinearity* **10** 243

[36] Barrio R, Blesa F, Dena A and Serrano S 2011 *Comput. Math. Appl.* **62** 4140–50

[37] Heitzig J 2002 Mappings between distance sets or spaces *PhD Thesis* Universität Hannover

[38] Scheffer M *et al* 2012 *Science* **338** 344–8

[39] Klinshov V V, Nekorkin V I and Kurths J 2015 *New J. Phys.* **18** 013004

[40] Hellmann F, Schultz P, Grabow C, Heitzig J and Kurths J 2016 *Sci. Rep.* **6** 1–12

[41] Mitra C, Kurths J and Donner R V 2015 *Sci. Rep.* **5** 1–10

[42] Mitra C, Choudhary A, Sinha S, Kurths J and Donner R V 2017 *Phys. Rev.* E **95** 032317

[43] Clark M M 2011 *Transport Modeling for Environmental Engineers and Scientists* (New York: Wiley)

[44] Cvitanović P, Artuso R, Mainieri R, Tanner G and Vattay G 2016 *Chaos: Classical and Quantum* (Copenhagen: Niels Bohr Inst.)

[45] Havlin S *et al* 2012 *Eur. Phys. J. Spec. Top.* **214** 273–93

[46] Giesl P and Hafstein S 2015 *Discrete Continuous Dyn. Syst.* B **20** 2291–331

[47] Van Rossum G and Drake F L Jr 1995 Python Reference Manual (Centrum voor Wiskunde en Informatica Amsterdam)

[48] Ascher D *et al* 2001 *Numerical Python*

[49] Jones E *et al* 2001 SciPy: Open Source Scientific Tools for Python

From Math to Metaphors and Back Again

Social-Ecological Resilience from a Multi-Agent-Environment Perspective

Jonathan F. Donges, Wolfram Barfuss

Science and policy stand to benefit from reconnecting the many notions of social-ecological resilience to their roots in complexity sciences. We propose several ways of moving towards operationalization through the classification of modern concepts of resilience based on a multi-agent-environment perspective.

From Math to Metaphors and Back Again.
Social-Ecological Resilience from a
Multi-Agent-Environment Perspective
GAIA 26/S1 (2017): 182–190

Abstract

Social-ecological resilience underlies popular sustainability concepts that have been influential in formulating the *United Nations Sustainable Development Goals (SDGs)*, such as the *Planetary Boundaries* and *Doughnut Economics*. Scientific investigation of these concepts is supported by mathematical models of planetary biophysical and societal dynamics, both of which call for operational measures of resilience. However, current quantitative descriptions tend to be restricted to the foundational form of the concept: persistence resilience. We propose a classification of modern notions of social-ecological resilience from a multi-agent-environment perspective. This aims at operationalization in a complex systems framework, including the persistence, adaptation and transformation aspects of resilience, normativity related to desirable system function, first- vs. second-order and specific vs. general resilience. For example, we discuss the use of the *Topology of Sustainable Management Framework*. Developing the mathematics of resilience along these lines would not only make social-ecological resilience more applicable to data and models, but could also conceptually advance resilience thinking.

Keywords

complex systems perspective, mathematical operationalization, multi-agent-environment systems, planetary boundaries, safe operating space for humanity, social-ecological resilience

Social-ecological system (SES) resilience is a popular concept now widely applied in many fields of science related to sustainable development as well as in science communication and education efforts (Folke et al. 2016, Folke 2016). Notably, the concept of resilience is at the heart of the *Planetary Boundaries Framework* (Rockström et al. 2009, Steffen et al. 2015), which, together with its extensions such as *Doughnut Economics* introducing the safe and just operating space for humanity (Raworth 2012), has been influential in formulating the *United Nations Sustainable Development Goals (SDGs)*[1]. However, as already Carpenter et al. (2001, p. 765) have pointed out: "Resilience has multiple levels of meaning: as a metaphor related to sustainability, as a property of dynamic models, and as a measurable quantity that can be assessed in field studies of SES". This multi-level nature of resilience can be seen as an intrinsic strength of the concept (e.g., Folke et al. 2016), but together with its often meandering use by various communities also has the potential to cause confusion and difficulties in operationalizing and practically applying the concept. The intention of this paper is to propose a classification of various modern concepts of social-ecological resilience from a multi-agent-environment perspective and, while not proposing a concrete operationalization, to discuss possible avenues to developing such a mathematical formalization reconnecting these notions to their theoretical foundations in complex systems theory.

SES resilience (Berkes and Folke 1998) originated from a complex systems perspective on ecological dynamics (Holling 1973) integrating at the time revolutionary mathematical insights into

[1] *www.un.org/sustainabledevelopment/sustainable-development-goals*

Contact: *Dr. Jonathan F. Donges* | Stockholm University | Stockholm Resilience Centre | Stockholm | Sweden | Tel.: +49 331 2882468 | E-Mail: donges@pik-potsdam.de

Wolfram Barfuss, M. Sc. | Humboldt-Universität zu Berlin | Department of Physics | Berlin | Germany | E-Mail: barfuss@pik-potsdam.de

both: Potsdam Institute for Climate Impact Research (PIK) | Research Domain for Earth System Analysis | Telegrafenberg A31 | 14473 Potsdam | Germany

https://doi.org/10.14512/gaia.26.S1.5

the properties of even relatively simple dynamical systems including nonlinearity, multistability, bifurcations and chaos (Lorenz 1963). From these insights, the basal understanding of resilience can be summarized as “the magnitude of disturbance that can be tolerated before a socioecological system (SES) moves to a different region of state space controlled by a different set of processes” (Carpenter et al. 2001, p. 765).

This classical definition of resilience resonated well beyond the area of theoretical research and translated into a concept of practical value for policy makers and participatory research endeavours. Thus, “more liberal definition(s)” of resilience emerged in this context such as the “capacity of a system to absorb disturbance and reorganize while undergoing change so as to still retain essentially the same function, structure, identity, and feedbacks“ (Scheffer 2009, p. 357). Eventually Folke et al. (2010) termed the integrated perspective of persistence, adaptation and transformation as “resilience thinking” based on Walker et al.’s (2004) seminal introduction of this triad of terms. This framework includes the spectrum from specific resilience of “what to what” (Carpenter et al. 2001) to general resilience (Carpenter et al. 2012). Additionally, Schneider and Vogt (2017, p. 179, in this issue) enrich this picture by distinguishing resilience of first-order associated to a specific system or actor from resilience of second-order that additionally encompasses the interactions of first-order resiliences of multiple systems or actors.

These extended definitions of SES resilience tend to use complex systems language metaphorically rather than focussing on operational measures and mathematical understanding. The purpose of this contribution is to argue for reconnecting these resilience metaphors to their foundations in complex systems theory.

>

EXHIBIT *SURVIVING THE FUTURE – RESILIENCE & DESIGN* (2016)

MAGIC SEVEN: “HOW CAN DESIGNERS BE SUPPORTED IN DEVELOPING CONTEMPORARY, RESILIENT DESIGNS?”

Resilience factors account for our ability to react to unforeseeable developments. For the project Magic Seven, *these were transferred to design: a creative range of questions was devised to inspire designers to think outside the box, think for the long haul, and incorporate incalculable courses of action into their designs. In the exhibition, the seven principles (adaptiveness, fault tolerance, modularity, longevity, assumption of unpredictability, diversity, self-learning ability) were illustrated by seven forks. The three-headed fork in the front center of the picture symbolizes alternative courses of action and stands for the diversity principle.*

© Monnier Ostermair

Reproduced from: J. F. Donges and W. Barfuss, From Math to Metaphors and Back Again Social-Ecological Resilience from a Multi-Agent-Environment Perspective Jonathan, GAIA, vol. 26, pp. 182-190, 2017, doi: 10.14512/gaia.26.S1.5. Published under Creative Commons Attribution License 3.0 (CC BY).

We believe that this agenda will serve to streamline communication on resilience across disciplines, help to avoid misunderstandings and improve the applicability of SES resilience concepts. In perspective, it will allow for devising useful quantitative measures capturing also more subtle aspects of SES resilience that are important for empirical measurements and applications to computer simulation models of SES across scales, for example, for use in advising policy makers. Beyond arguing for these more practical benefits of quantification and formalization, we follow the reasoning of Carpenter et al. (2001, p. 767) that a theory's "success is measured by the utility of the concepts in terms of their ability to influence the research topics chosen by scientists and stimulate productive hypotheses", and "progress in the definition of concepts is central to advancement of science".

Persistence Resilience: Rooted in a Complex Systems Perspective

The persistence aspect of SES resilience is the most formalized among the various other notions such as adaptation and transformation resilience. It corresponds to the foundational dynamical systems understanding of *ecological resilience* (Holling 1973): "the magnitude of disturbance a system can tolerate before it moves into a different (region of) state (space)" (Scheffer 2009, p. 357). In this view, the state of a system is formally described by a set of state variables (see dark green axes in figure 1), where the state at a particular time corresponds to a point or state vector in a potentially high-dimensional state space. The system state evolves in time along a trajectory following prescribed deterministic or stochastic rules.

In what are often called complex dynamical systems, multiple *attractors* can coexist in state space implying multistability, that is, the system can evolve towards alternative attractors depending on in which *basin of attraction* the initial system state lies. For example, in the domain of ecological resilience, turbid and clear attracting states of a lake can coexist in state space (Scheffer 2009). This property of multistability is central to formal definitions of persistence resilience and is captured visually by the so-called ball-and-cup diagram (figure 1a). The ball symbolizes the current system state. The minima of the stability landscape correspond to fixed point attractors. In analogy to a ball rolling along a hilly landscape, the cups or valleys depict the attractors' basins of attraction.

Generally, mathematical descriptions of persistence resilience build upon this picture of a dynamical system evolving in a state space with multiple attractors. A perturbation, shock or disturbance is then often seen as a sudden shift of the system state away from dynamical equilibrium (i.e., with the system residing on an attractor) induced by some external force. Measures of persistence resilience can then be related to various dynamical and geometrical properties of the attractor and its basin of attraction. Among others, operational measures of persistence resilience can be derived from the speed of return to the attractor after small perturbations (so-called *engineering resilience* related to linear stability concepts in dynamical systems theory, see Pimm 1984, Anderies et al. 2013), the attractors' distance to its basin boundary (Klinshov et al. 2015), the volume of the basin of attraction (Scheffer et al. 2001, Menck et al. 2013), or combinations thereof (Mitra et al. 2015, Hellmann et al. 2016). Recent work on early warning signals for critical transitions in SES (Scheffer et al. 2009) exemplifies a fruitful application of these and related mathematical formalizations of persistence resilience.

Resilience Thinking: Modern Concepts of Social-Ecological Resilience

While the formal study of persistence resilience is quite elaborate, it has been recognized that accounting for persistence aspects is not sufficient in a complex, nonlinear world. Walker et al. (2004) extended the persistence notion of resilience with the aspects of *adaptation* and *transformation*.

Adaptability usually refers to the capacity of a system to learn and adjust its responses to changing external processes within the current stability domain (Berkes et al. 2003); put in short "to manage (persistence) resilience" (Walker et al. 2004). An important extension in the mental model has been made at this point. Whereas persistence resilience can be defined in a dynamical systems model, the notion of adaptability requires thinking additionally of an agent, an entity capable of choosing among a certain set of actions. The distinction between the persistence and adaptation aspects of resilience has been reflected already through the *adaptive cycle* concept (Gunderson 2001) and in the seminal work by Holling (1973). A view going beyond this notion describes adaptability as the ability to maintain system functioning under a changing environment (Martin-Breen and Anderies 2011). This definition allows the system to modify its current attractor and the associated basin of attraction as long as the functioning of the system is ensured. What system function is considered as desirable here needs to be specified in addition. This is a normative notion that needs to be accounted for in advanced complex systems operationalizations of resilience to be outlined in the next section. Similarly, the resilience of "what to what" (Carpenter et al. 2001) has to be specified, for example, the resilience of a certain system property or function for a certain attractor with respect to a specific (fast) change of system state (shock) that may be either unforeseeable or anticipable. Also (slow) changes in the functioning and dynamics of the environment are possible influences a system can be resilient against (via adaptation).

Transformability recognizes that even an adaptation view of SES resilience is not sufficient and refers to the "capacity to create a fundamentally new system when ecological, economic, or social conditions make the existing system untenable" (Walker et al. 2004). Along these lines, the notion of *general resilience* acknowledges the fact that building *specific resilience* for one part of the system does not guarantee increasing specific resilience in other parts or the whole system, or may even undermine general resilience of the whole system. It therefore acknowledges the dangers of a too narrow perspective, for example, focussing only on the specific resilience of social or ecological subsystems of an SES (Carpenter et al. 2012). Both recognize SES as complex adaptive systems (Martin-Breen and Anderies 2011, Folke 2016). However, it remains unclear what makes a system *fundamentally* new and what is the exact difference between adaptation and transformation.

In summary, while persistence resilience is founded on deterministic concepts from the theory of dynamical systems (ultimately going back to Newton's classical mechanics), modern notions of SES resilience such as those related to adaptability and transformability, specific vs. general and first- vs. second-order resilience at their core require introducing agency into efforts towards much needed mathematical operationalizations. In the subsequent section, we contribute to this endeavour by classifying and discussing modern resilience notions from a multi-agent-environment perspective.

Notions of Social-Ecological Resilience from a Multi-Agent-Environment Perspective

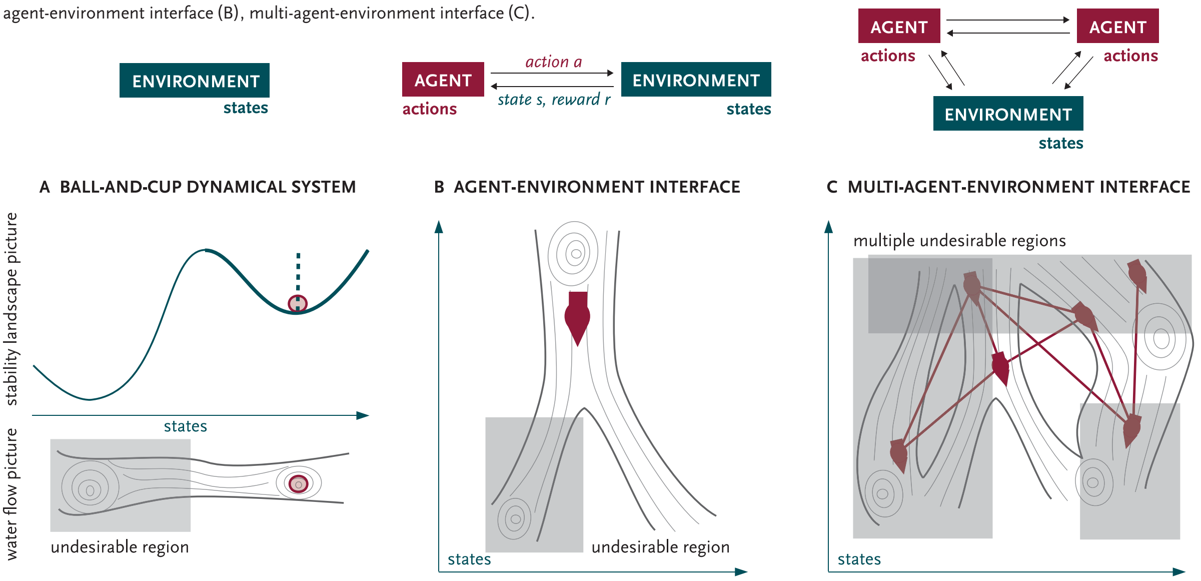
In the following we outline how the resilience triad of persistence, adaptation and transformation (Folke et al. 2010) could be mathematically operationalized on the foundation of multi-agent environment systems that are well established in computer science (Busoniu et al. 2008) and that show parallels to Ostrom's conceptualization of SES (Ostrom 2009). We discuss normative notions related to the desirability of system states and classifications such as specific vs. general and first- vs. second-order resilience. However, we stress that it is beyond the scope of this article to fully develop the proposed agenda and that the following discussion outlines only one of potentially many possible operationalizations.

We propose three levels of SES resilience complexity (figure 1). The first level focuses on the persistence aspect described in the previous section (figure 1a). The term *environment* denotes the ecological, social and economic stochastic or deterministic system dynamics without any agent behavior. The system function notion of persistence resilience is connected to the desirability of system states: the gray area in figure 1a indicates states that are perceived as undesirable.

To describe the adaptation and transformation aspects of SES resilience, a "ball" representing the system state alone is not sufficient. Instead, moving to a second level of resilience complexity by introducing an agent equipped with the agency to choose among a set of actions is required (figure 1b). Schellnhuber (1998, 1999) already introduced related ideas in the Earth system context under the terms *geocybernetics* and *Earth system analysis* distinguishing the *ecosphere* and *anthroposphere* (together constituting the environment) from the *global subject* (the agent). Similarly, Anderies et al. (2007) – inspired by Ostrom's general framework to study SES (Ostrom 2009) – take a single-agent-environment perspective to study SES following a robustness approach. Introducing an agent extends the environment of the persistence resilience case to an agent-environment interface and simultaneously to a decision problem of what action to choose given a history of system observations. Any decision-making framework requires stating the choices or actions available to the decision maker and a criterion to evaluate the decisions, often called either rewards, utility or costs associated with the actions (Steele and Stefánsson 2016). With the agent's

>

FIGURE 1: Three levels of increasing resilience complexity: dynamical system (environment) (A), agent-environment interface (B), multi-agent-environment interface (C).



Reproduced from: J. F. Donges and W. Barfuss, From Math to Metaphors and Back Again Social-Ecological Resilience from a Multi-Agent-Environment Perspective Jonathan, GAIA, vol. 26, pp. 182-190, 2017, doi: 10.14512/gaia.26.S1.5. Published under Creative Commons Attribution License 3.0 (CC BY).

strategy or policy we refer to the rule describing what action to apply given a history of observations. Fixing a (default) strategy, this system can be described equivalently to the dynamical systems case (first level of resilience complexity, figure 1a), that is, accounting for persistence resilience is also applicable here. This is visualized by the "default" flow in figure 1b. Hence, a change of strategy is equivalent to a change of the stability landscape in the ball-and-cup picture. To see the correspondence of the latter and the agent-environment view, imagine the agent climbing a hill (i.e., applying management deviating from the default strategy) in the stability landscape along a specific direction in state space. The agent's movement is equally well described by a different landscape in the ball-and-cup picture, in which the ball (now representing the agent) glides downhill in the same direction following the default flow.

Introducing an agent allows us to consider *meta-rules* or *algorithms* that govern how a strategy adjusts to the environment over time. These meta-rules may be inspired by modern artificial intelligence or machine learning algorithms (Sutton and Barto 1998) combined with *sustainability paradigms* as proposed by Schellnhuber (1998, 1999): optimization, pessimization, equitization, or standardization. For example, the equitization paradigm bears the maxim that the option space for future generations is kept as open as possible by actions of the current generation for building resilience (see also Vogt 2013). Practically this requires suitable meta-rules to govern multiple kinds of uncertainties and risks (Renn 2008).

In terms of desirability, at least two options seem plausible: 1. one can either fix the desirability of a state, or 2. the evaluation criterion of the decision context can be utilized. In the former case, a state's desirability is independent from the current strategy, whereas in the latter case the desirability of a state results from the reward the agents receive following that current strategy. The *Tolerable Windows Approach* (Petschel-Held et al. 1999) and the *Planetary Boundaries Framework* (Rockström et al. 2009, Steffen et al. 2015) are examples of a division of state space into desirable and undesirable states in the sustainability context. While the desirability of states depends on normative judgement, this does not necessarily hold for SES resilience, since an undesirable state may be resilient as well (Carpenter et al. 2001).

The third level of resilience complexity extends the agent-environment further to a multi-agent-environment system (figure 1c, Busoniu et al. 2008). While all characteristics of the agent-environment interface discussed above apply, the multi-agent aspect allows for the possibility of emergent phenomena (Sawyer 2005). It further emphasizes the potentially conflicting interests of the agents, visualized by the distinct individual desirability regions in state space.

In the following we discuss how some of the modern notions of social-ecological resilience integrate into the proposed three levels of resilience complexity (see table 1 for an overview).

TABLE 1: Applicability of resilience concepts (rows) to our proposed levels of resilience complexity (columns).

	ENVIRONMENT	AGENT-ENVIRONMENT	MULTI-AGENT-ENVIRONMENT
type	persistence	persistence adaptation transformation	persistence adaptation transformation
scale		first-order	first-order second-order
scope	specific general	specific general	specific general

Adaptation and transformation. Folke et al. (2010, p. 2) describe adaptability as the "capacity of a SES to learn, combine experience and knowledge, adjust its responses to changing external drivers and internal processes, and continue developing within the current stability domain or basin of attraction". Transformability is described as the "capacity to transform the stability landscape itself". In our view both aspects can only be treated either in the second or third level of resilience complexity: the agent-environment or the multi-agent-environment case, in which the agents use an internal meta-rule or algorithm to derive the actual rule (strategy) describing what action to apply given a history of observations. Typically these algorithms are constantly changing their internal variables, representing implicitly the agents' world model, as a reaction to the observations over time.

Interpreting these definitions of adaptation and transformation in their most narrow sense, any change of the internal variables of the meta-rule or learning algorithm is an adaptation, as long as it does not change the actual strategy. If the strategy changes one has to speak of a transformation, essentially because a change of strategy is equivalently describable by a change of the stability landscape.

As an alternative interpretation one may include into adaptations changes in the strategies altering the corresponding stability landscape smoothly, that is, those changes that vary the shape (e.g., height, extent) or location of the minimum without disrupting the structure of this landscape. In contrast, a transformation could be defined as strategy changes that alter the stability landscape qualitatively: destruction of old or creation of new attractors. Technically, these situations are commonly referred to as bifurcations, tipping points or critical transitions (Scheffer 2009). This interpretation focuses on the fact that a transformation is perceived as a "fundamental" change (Walker et al. 2004, Folke et al. 2010).

A further distinction between adaptation and transformation could build on the dialectic micro-macro relationship between an agent and the social structure connecting agents in the multi-agent-environment perspective. An adaptation would correspond to strategy changes of an individual that do not alter a suitable macroscopic description of the multi-agent system including the complex network structure of social-ecological interactions, whereas transformations are observable qualitatively on the macroscopic level (Lade et al. 2017). As a simple example, one microscopic variable could be the wealth of an agent, while the macroscopic observable is average wealth.

Specific and general resilience. Specific resilience refers to resilience of "what to what" (Carpenter et al. 2001) whereas general resilience is described as "the capacity of social-ecological systems to adapt or transform in response to unfamiliar or unknown shocks" (Carpenter et al. 2012, p. 3251). We here interpret specific resilience as the capacity to absorb shocks along a specific dimension of the state space (or a more general subset of dimensions) including fast (states) and slow variables (parameters). For example, while we illustrated the persistence aspect with only one dimension in state space (figure 1a), the agent-environment interfaces (figure 1b,c) are visualized with two dimensions. Depending on context, both projections may be radical simplifications of the actual high dimensional state space. Building resilience for a *specific* subset of these state space dimensions could correspond to increasing the basin of attraction only along these dimensions.

General resilience, however, acknowledges the importance of the total size of the basin of attraction, that is, where the direction of the shocks is not specified. Further it takes into account the interactions between different state space dimensions, that is, whether the increase of the basin of attraction in one dimension may change the basin's size in other dimensions. With this interpretation one can distinguish specific and general resilience in all levels of resilience complexity presented in figure 1 (see also table 1).

Resilience of first and second order. In analogy to the distinction of specific and general resilience, Schneider and Vogt (2017, p. 179, in this issue) discuss the notions of resilience of first and second order. They define resilience of first order for a specific system, entity, institution, or actor. Resilience of second order takes a perspective to include the relationships of a specific entity to further actors, structures, and contexts. We interpret the focus of the concept of resilience *order* to be the actor or agent. Thus we suggest formalizing resilience of first order as the resilience associated with one agent (figure 1b). Correspondingly, the resilience of second order asks how the resilience of one agent affects the resiliences of other agents in a multi-agent-environment system (see figure 1c). Thus, building resilience of second order demands building resilience of first order of individual agents in a mutually beneficial way. It is an interesting research question to ask what properties of the meta-rules or adaptation algorithms are required for this interpretation of second-order resilience.

Example: *Topology of Sustainable Management Framework*

To give a concrete example how our proposed classification of resilience complexities can bring new insights by a rigorous mathematical treatment, we briefly introduce the *Topology of Sustainable Management Framework* (Heitzig et al. 2016). Extending upon

>

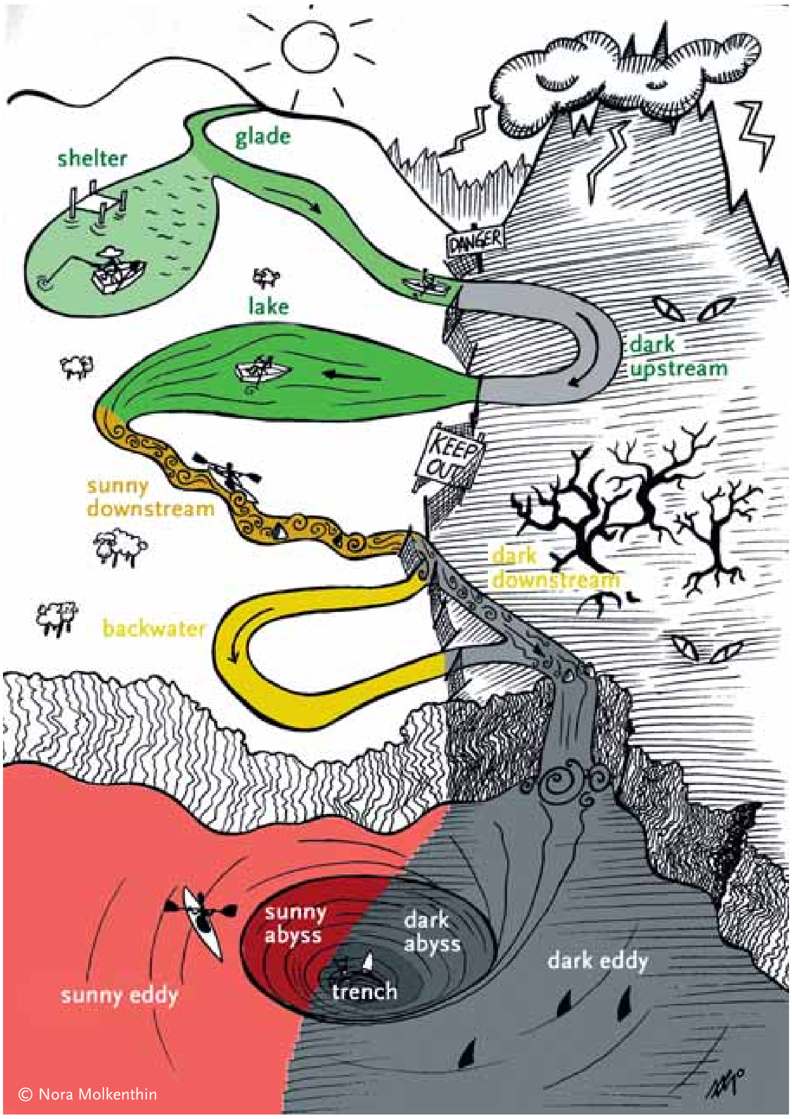


FIGURE 2: Illustration of the mathematical *Topology of Sustainable Management Framework* formalizing resilience based on an agent-environment perspective (modified from Heitzig et al. 2016).

related efforts to formalize resilience using viability theory (Deffuant and Gilbert 2011), Heitzig et al. (2016) show how a classification of qualitatively different regions in system state space emerges from the following three ingredients: 1. environmental dynamics under a default strategy, 2. available management options the agent can choose from, and 3. the division of state space into desirable ("sunny") and undesirable ("dark") regions. Hence, it uses an agent-environment interface perspective (second level of resilience complexity) with a default strategy and fixed state desirability (figure 2). The various elements of this picture metaphorically illustrate the underlying mathematical *Topology of Sustainable Management Framework* with the waterstream corresponding to the stability landscape under the default policy, similarly as in figure 1. The interested reader is referred to the original publication for the mathematical details.

For example, the *shelter* is the sunny set of states in which the agent can remain forever without any management. Both in the *glade* and the *lake* it is possible to reach the shelter but the agent has to apply management. From the lake it has to cross through the dark region, whereas from the glade it can reach the shelter without leaving the sunny region. In other regions, such as the *backwaters,* the shelter cannot be reached, but the agent can remain in the sunny region by constant or repeated management. These regions emerge from the allowed rule changes describing how the agent is able to adapt to and manage the environment.

In the *Topology of Sustainable Management Framework,* the default action is perceived as preferable to the other available management options. The rationale is that non-default actions are at risk of becoming inoperative, for example, due to external shocks. Hence, the default is considered as a safer option. Several dilemmas arise from this reasoning: for example, starting in the lake the agent can either remain in the sunny region under constant and potentially risky management or choose to cross the dark region to reach the shelter, where no management is needed. Note that while the mathematical framework serves to highlight these dilemmas, resolving them requires deep ethical considerations taking into account questions of justice, freedom and identity (Vogt 2013). See Heitzig et al. (2016) for a discussion of further dilemmas and various example systems to which the framework has been applied.

Generalizing from measures of persistence resilience discussed above, characteristics of these regions such as volume, depth, distance from the boundary or return rate could be interpreted as a sequence of operational measures capturing both persistence and adaptation aspects of SES resilience. For example, *shelter resilience* could correspond to the volume of the shelter region and indicate the size of a shock the system is capable to absorb to remain in the shelter without using management. Moreover, assuming the system continues to reside in the shelter or glade, *glade resilience* could correspond to the size of the glade plus the size of the shelter. This measure would indicate the magnitude of shock the system is capable to absorb under the potential need to apply management to return to the shelter without leaving the sunny region. Note that in this particular example all adaptation and transformation aspects of resilience have been incorporated into the classification of state space by emphasis on the default policy flow and possible management options deviating from the default action.

The point we would like to make with these examples is that even from fairly simple ingredients a rich and sometimes unintuitive picture of resilience may emerge under formal treatment with broad potential for applications. Future mathematical work could further extend the *Topology of Sustainable Management Framework* to accommodate more advanced resilience dimensions such as specific vs. general and first- vs. second-order resilience, for example, by including multiple agents with possibly distinct management options and desirability criteria.

> *Reconnecting modern concepts of social-ecological resilience with their roots in complex systems theory, based on a multi-agent-environment perspective, is relevant for analytically addressing global change problems of the Anthropocene.*

Discussion: Earth System Resilience in the Anthropocene

Reconnecting modern concepts of social-ecological resilience with their roots in complex systems theory is relevant for analytically addressing global change problems of the Anthropocene (Haber et al. 2016). Viewed from a resilience angle, the *SDGs* can be interpreted as normative criteria defining desirable biophysical, social and economic Earth system states (Folke et al. 2016). The *Planetary Boundaries* (Rockström et al. 2009, Steffen et al. 2015) and related concepts such as *Doughnut Economics* (Raworth 2012) argue for biosphere stewardship to maintain a safe (and just) operating space (Vogt 2013, Ekardt 2016). This is where the planetary SES is seen as resilient and where development towards the *SDGs* is argued to be possible. Refined insights into principles for actively building resilience and their preconditions would be useful in this context. To this end, Biggs et al. (2015) summarize seven principles for building resilience including its persistence, adaptation, and transformation aspects: 1. maintain diversity and redundancy, 2. manage connectivity, 3. manage slow variables and feedbacks, 4. foster complex adaptive systems thinking, 5. encourage learning, 6. broaden participation, and 7. promote polycentric governance. These principles for building resilience can also be viewed in their inverse forms as principles for undermining resilience of undesirable system states and structures.

Operational measures of the various dimensions of resilience as outlined above including persistence, adaptation, transformation, first- vs. second-order and specific vs. general resilience could be employed for systematically evaluating these seven and more principles for building (or undermining) resilience and their detailed preconditions. Such an investigation would be supported by computer simulation models of SES of interest (Schlüter et al. 2012) but could also integrate various sources of empirical data. Using this approach, the validity of the principles for building resilience can be assessed in different situations, including possible unintended side effects induced by applying them.

Most analytical studies on the resilience of SES and the associated principles for building resilience have been conducted on local and regional scales. But a key characteristic of the Anthropocene and the inherent great social and environmental challenges are ever densifying global networks of teleconnected and tightly intertwined social-ecological processes. Therefore, computer simulation models as well as more stylized conceptual models are needed to operationally study resilience and principles for building resilience for the planetary SES that capture coevolving and networked biophysical, socio-economic and socio-cultural dynamics (Verburg et al. 2016, Donges et al. 2017). Applied in this setting, operational measures of resilience dimensions will serve as valuable tools for Earth system analysis (Schellnhuber et al. 2004).

As a recent example, it has been argued that to meet the Paris climate agreement (UNFCCC 2015) a controlled collapse of the planetary-scale fossil fuel sector must be induced for triggering a rapid global decarbonization transformation (Schellnhuber et al. 2016) as part of a concerted broader sustainability transformation (WBGU 2011). From a scientific perspective, this agenda calls for a deeper understanding of the apparently massive specific resilience of this part of the global SES, the associated general planetary SES resilience and principles for undermining this specific resilience without harmful and unwanted side effects such as economic crises. Reconnecting modern concepts from resilience thinking to their formal complex systems foundations through a multi-agent-environment perspective as proposed in this article could make a useful contribution to this endeavour by providing operational measures of various resilience dimensions and, more fundamentally, by shedding light on the underlying structure of modern resilience concepts and their interconnections.

This work was conducted in the framework of the project on *Coevolutionary Pathways in the Earth System (COPAN)* at the Potsdam Institute for Climate Impact Research (PIK). We thank *Boyan Beronov, Jobst Heitzig, Steven Lade, Wolfgang Lucht* and *Finn Müller-Hansen* for inspiring discussions on the concept of resilience. We are grateful for financial support by the Stordalen Foundation (via the *Planetary Boundary Research Network PB.net*), the *Earth League's EarthDoc Programme*, Leibniz Association (project *DOMINOES*) and the Heinrich Böll Foundation.

References

Anderies, J. M., C. Folke, B. Walker, E. Ostrom. 2013. Aligning key concepts for global change policy: Robustness, resilience, and sustainability. *Ecology and Society* 18/2: 8.

Anderies, J. M., A. A. Rodriguez, M. A. Janssen, O. Cifdaloz. 2007. Panaceas, uncertainty, and the robust control framework in sustainability science. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 104/39: 15194–15199.

Berkes, F., C. Folke. 1998. *Linking social and ecological systems for resilience and stability.* Cambridge, UK: Cambridge University Press.

Berkes, F., J. Colding, C. Folke. 2003. *Navigating social-ecological systems: Building resilience for complexity and change.* Cambridge, UK: Cambridge University Press.

Biggs, R., M. Schlüter, M. L. Schoon (Eds.). 2015. *Principles for building resilience: Sustaining ecosystem services in social-ecological systems.* Cambridge, UK: Cambridge University Press.

Busoniu, L., R. Babuska, B. de Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)* 38/2: 156.

Carpenter, S. R., B. Walker, J. M. Anderies, N. Abel. 2001. From metaphor to measurement: Resilience of what to what? *Ecosystems* 4/8: 765–781.

Carpenter, S. R. et al. 2012. General resilience to cope with extreme events. *Sustainability* 4/12: 3248–3259.

Deffuant, G., N. Gilbert (Eds.). 2011. *Viability and resilience of complex systems: Concepts, methods and case studies from ecology and society.* Berlin: Springer.

Donges, J. F., W. Lucht, F. Müller-Hansen, W. Steffen. 2017. The technosphere in Earth system analysis: A co-evolutionary approach. *Anthropocene Review* 4/1: 23–33.

Ekardt, F. 2016. *Theorie der Nachhaltigkeit: Ethische, rechtliche, politische und transformative Zugänge – am Beispiel von Klimawandel, Ressourcenknappheit und Welthandel.* 3rd edition. Baden-Baden: Nomos.

Folke, C. 2016. Resilience. In: *Oxford Research Encyclopedia of Environmental Science.* doi: 10.1093/acrefore/9780199389414.013.8.

Folke, C., R. Biggs, A. Norström, B. Reyers, J. Rockström. 2016. Social-ecological resilience and biosphere-based sustainability science. *Ecology and Society* 21/3: 41.

Folke, C., S. R. Carpenter, B. Walker, M. Scheffer, T. Chapin, J. Rockström. 2010. Resilience thinking: Integrating resilience, adaptability and transformability. *Ecology and Society* 15/4: 20.

Gunderson, L. H. 2001. *Panarchy: Understanding transformations in human and natural systems.* Washington, D.C.: Island Press.

Haber, W., M. Held, M. Vogt (Eds.). 2016. *Die Welt im Anthropozän. Erkundungen im Spannungsfeld zwischen Ökologie und Humanität.* Munich: oekom.

Heitzig, J., T. Kittel, J. F. Donges, N. Molkenthin. 2016. Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system. *Earth System Dynamics* 7/1: 21–50.

Hellmann, F., P. Schultz, C. Grabow, J. Heitzig, J. Kurths. 2016. Survivability of deterministic dynamical systems. *Scientific Reports* 6: 29654.

Holling, C. S. 1973. Resilience and stability of ecological systems. *Annual Review of Ecology and Systematics* 4: 1–23.

Klinshov, V. V., V. I. Nekorkin, J. Kurths. 2015. Stability threshold approach for complex dynamical systems. *New Journal of Physics* 18/1: 013004.

Lade, S. et al. 2017. *Modelling social-ecological transformations: An adaptive network proposal.* arXiv:1704.06135 [nlin.AO] (accessed June 27, 2017).

>

Lorenz, E. N. 1963. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences* 20/2: 130–141.

Martin-Breen, P., J. M. Anderies. 2011. *Resilience: A literature review.* http://opendocs.ids.ac.uk/opendocs/handle/123456789/3692 (accessed June 27, 2017).

Menck, P. J., J. Heitzig, N. Marwan, J. Kurths. 2013. How basin stability complements the linear-stability paradigm. *Nature Physics* 9/2: 89–92.

Mitra, C., J. Kurths, R. V. Donner. 2015. An integrative quantifier of multistability in complex systems based on ecological resilience. *Scientific Reports* 5: 16196.

Ostrom, E. 2009. A general framework for analyzing sustainability of social-ecological systems. *Science* 325/5939: 419–422.

Petschel-Held, G., H. J. Schellnhuber, T. Bruckner, F. L. Toth, K. Hasselmann. 1999. The tolerable windows approach: Theoretical and methodological foundations. *Climatic Change* 41/3–4: 303–331.

Pimm, S. L. 1984. The complexity and stability of ecosystems. *Nature* 307/5949: 321–326.

Raworth, K. 2012. A safe and just space for humanity: Can we live within the doughnut. *Oxfam Policy and Practice: Climate Change and Resilience* 8/1: 1–26.

Renn, O. 2008. *Risk governance: Coping with uncertainty in a complex world.* Milton Park, UK: Earthscan.

Rockström, J. et al. 2009. A safe operating space for humanity. *Nature* 461/7263: 472–475.

Sawyer, R. K. 2005. *Social emergence: Societies as complex systems.* New York: Cambridge University Press.

Scheffer, M. 2009. *Critical transitions in nature and society.* Princeton, NJ: Princeton University Press.

Scheffer, M., S. R. Carpenter, J. A. Foley, C. Folke, B. Walker. 2001. Catastrophic shifts in ecosystems. *Nature* 413/6856: 591–596.

Scheffer, M. et al. 2009. Early-warning signals for critical transitions. *Nature* 461/7260: 53–59.

Schellnhuber, H. J. 1998. Discourse: Earth system analysis – The scope of the challenge. In: *Earth System Analysis.* Edited by H. J. Schellnhuber, V. Wenzel. Berlin: Springer. 3–195.

Schellnhuber, H. J. 1999. "Earth system" analysis and the second Copernican revolution. *Nature* 402: C19–C23.

Schellnhuber, H. J. et al. 2004. *Earth system analysis for sustainability.* Cambridge, MA: MIT Press.

Schellnhuber, H. J., S. Rahmstorf, R. Winkelmann. 2016. Why the right climate target was agreed in Paris. *Nature Climate Change* 6/7: 649–653.

Schlüter, M. et al. 2012. New horizons for managing the environment: A review of coupled social-ecological systems modeling. *Natural Resource Modeling* 25/1: 219–272.

Schneider, M., M. Vogt. 2017. *Responsible resilience:* Rekonstruktion der Normativität von Resilienz auf Basis einer responsiven Ethik. *GAIA* 26/S1: 174–181.

Steele, K., H. Stefánsson. 2016. Decision Theory. In: *The Stanford Encyclopedia of Philosophy.* Edited by E. N. Zalta. https://plato.stanford.edu/archives/win2016/entries/decision-theory (accessed June 27, 2017).

Steffen, W. et al. 2015. Planetary boundaries: Guiding human development on a changing planet. *Science* 347/6223: 1259855.

Sutton, R. S., A. G. Barto. 1998. *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

UNFCCC (United Nations Framework Convention on Climate Change). 2015. *Adoption of the Paris Agreement.* FCCC/CP/2015/L.9/Rev.1.

Verburg, P. H. et al. 2016. Methods and approaches to modelling the Anthropocene. *Global Environmental Change* 39: 328–340.

Vogt, M. 2013. *Prinzip Nachhaltigkeit: ein Entwurf aus theologisch-ethischer Perspektive.* 3rd edition. Munich: oekom.

Walker, B., C. S. Holling, S. R. Carpenter, A. Kinzig. 2004. Resilience, adaptability and transformability in social-ecological systems. *Ecology and Society* 9/2: 5.

WBGU (German Advisory Council on Global Change). 2011. *World in transition – A social contract for sustainability.* Berlin: WBGU.

Submitted December 1, 2016; revised version accepted June 19, 2017.

Jonathan F. Donges

Born 1983 in Engelskirchen, Germany. 2012 PhD in theoretical physics at Humboldt-Universität zu Berlin. Since 2013 joint PostDoc at the Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany, and Stockholm Resilience Centre, Sweden. Co-leader of PIK's flagship project on *Coevolutionary Pathways in the Earth System (COPAN).* Research interests: models of planetary-scale social-ecological dynamics in the Anthropocene, Earth system and social tipping elements and their interactions, modelling transformative change in social-ecological systems focussing on sustainability transformation, concepts, and measures of resilience.

Wolfram Barfuss

Born 1990 in Fürth, Germany. 2015 M. Sc. Physics, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany. Since 2015 PhD studies at the Potsdam Institute for Climate Impact Research (PIK), Potsdam, Germany, and Humboldt-Universität zu Berlin. Member of the Heinrich Böll Foundation Cluster *Transformation.* Research interests: adaptive agent behavior in conceptual social-ecological system models.

GAiA ECOLOGICAL PERSPECTIVES FOR SCIENCE AND SOCIETY

© 2017 Verein Gaia | Konstanz, St. Gallen, Zurich

EDITOR-IN-CHIEF
Irmi Seidl | Birmensdorf
(responsible according to the press law)

EDITORIAL OFFICE
Dr. Almut Jödicke | ETH Zentrum |
PO Box CAB 42 | 8092 Zurich | Switzerland |
E-Mail: redgaia@env.ethz.ch

Dr. Martina Blum/Tobias Mickler | oekom verlag |
Waltherstr. 29 | 80337 Munich | Germany |
E-Mail: blum@oekom.de/mickler@oekom.de

Dr. Ulrike Sehy | oekom verlag | Hagenbuchrain 13 |
8047 Zurich | Switzerland | E-Mail: sehy@oekom.ch

GRAPHIC DESIGN + TYPESET
Heike Tiller | Munich | E-Mail: h.tiller@freenet.de

PUBLISHER
oekom verlag – Gesellschaft für ökologische Kommunikation mbH | Waltherstr. 29 | 80337 Munich | Germany | *www.oekom.de* | Partners and shareholders: *Jacob Radloff,* Feldafing, 77 percent, and *Christoph von Braun,* Munich, 23 percent

ADVERTISEMENTS
Janine Gaumer | oekom GmbH |
Tel.: +49 89 54418435 | E-Mail: anzeigen@oekom.de

PRINTER
Kessler Druck + Medien | 86399 Bobingen |
Germany | *www.kesslerdruck.de*

FREQUENCY Four times a year.

SUBSCRIPTION
Trial subscription (2 issues including shipping in Germany): 22.10 EUR | Subscription: private: 108.00 EUR; institutional: 195.00 EUR; reduced: 73.00 EUR | Single issue: 27.30 EUR. VAT included, plus shipping. Cancellations six weeks before end of subscription year. Payment in advance.

SUBSCRIPTIONS, DISTRIBUTION
InTime Media Services GmbH | Zeitschriften oekom |
PO Box 1363 | 82034 Deisenhofen | Germany |
Tel.: +49 89 85853570 | Fax: +49 89 8585362570 |
E-Mail: oekom@intime-media-services.de
www.oekom.de/gaia/abonnement

ACCOUNT for Germany: Postbank München |
IBAN DE59 7001 0080 0358 7448 03 | BIC PBNKDEFF |
for Switzerland: PostFinance SWISS POST |
IBAN CH73 0900 0000 4019 4607 4 | BIC POFICHBE

Since 2008, oekom offsets its unavoidable CO_2 emissions.

ClimatePartner

ISSN 0940-5550

Printed on
Circle Offset
Premium White,
certified with
The Blue Angel
(RAL-UZ 14).

3.2 *Sustainable management of complex social-ecological systems*

This section addresses the challenge of managing complex social-ecological systems so that they remain within the planetary boundaries [Rockström et al., 2009] and at the same time respect important social foundations [Raworth, 2017].

In the first paper, "Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the Earth system" [Heitzig et al., 2016], we developed a mathematical theory of the qualitative topology of sustainable management.

This framework was applied numerically in "From lakes and glades to viability algorithms: Automatic classification of system states according to the Topology of Sustainable Management" [Kittel et al., 2017b, not included in this reader] to a very stylized World-Earth model. Furthermore, we studied the difference between one-time versus permanent management.

An alternative approach to managing social-ecological systems is machine learning. In "Deep reinforcement learning in World-Earth system models to discover sustainable management strategies" [Strnad et al., 2019] we explored this methodology for finding sustainable management trajectories.

Proposals for management of the Earth system usually follow one of three policy paradigms: the paradigm of sustainability, the concept of the safe operating space or the optimization of economic welfare [Schellnhuber, 1998, Petschel-Held et al., 1999]. The study "When optimization for governing human-environment tipping elements is neither sustainable nor safe" [Barfuss et al., 2018] finds that none of these policy paradigms guarantees fulfilling requirements imposed by another paradigm and derive simple heuristics for the conditions under which these trade-offs occur. It is shown that the absence of such a master paradigm is of special relevance for governing real-world tipping systems such as climate, fisheries, and farming, which may reside in a parameter regime where economic optimization is neither sustainable nor safe.

Finally, we close with "A Thought Experiment on Sustainable Management of the Earth System" [Heitzig et al., 2018]. Here we analysed a possible management dilemma that humanity might be currently faced with in dealing with current environmental challenges of the Anthropocene and compared various theoretical approaches for analysing it.

Earth Syst. Dynam., 7, 21–50, 2016
www.earth-syst-dynam.net/7/21/2016/
doi:10.5194/esd-7-21-2016

Earth System Dynamics
Open Access
EGU

Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the Earth system

J. Heitzig[1], T. Kittel[1,2], J. F. Donges[1,3], and N. Molkenthin[4]

[1]Research Domains Transdisciplinary Concepts & Methods and Earth System Analysis, Potsdam Institute for Climate Impact Research, P.O. Box 60 12 13, 14412 Potsdam, Germany
[2]Department of Physics, Humboldt University, Newtonstr. 15, 12489 Berlin, Germany
[3]Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden
[4]Department for Nonlinear Dynamics & and Network Dynamics Group, Max Planck Institute for Dynamics and Self-Organization, Bunsenstraße 10, 37073 Göttingen, Germany

Correspondence to: J. Heitzig (heitzig@pik-potsdam.de)

Received: 13 February 2015 – Published in Earth Syst. Dynam. Discuss.: 12 March 2015
Revised: 9 November 2015 – Accepted: 26 November 2015 – Published: 18 January 2016

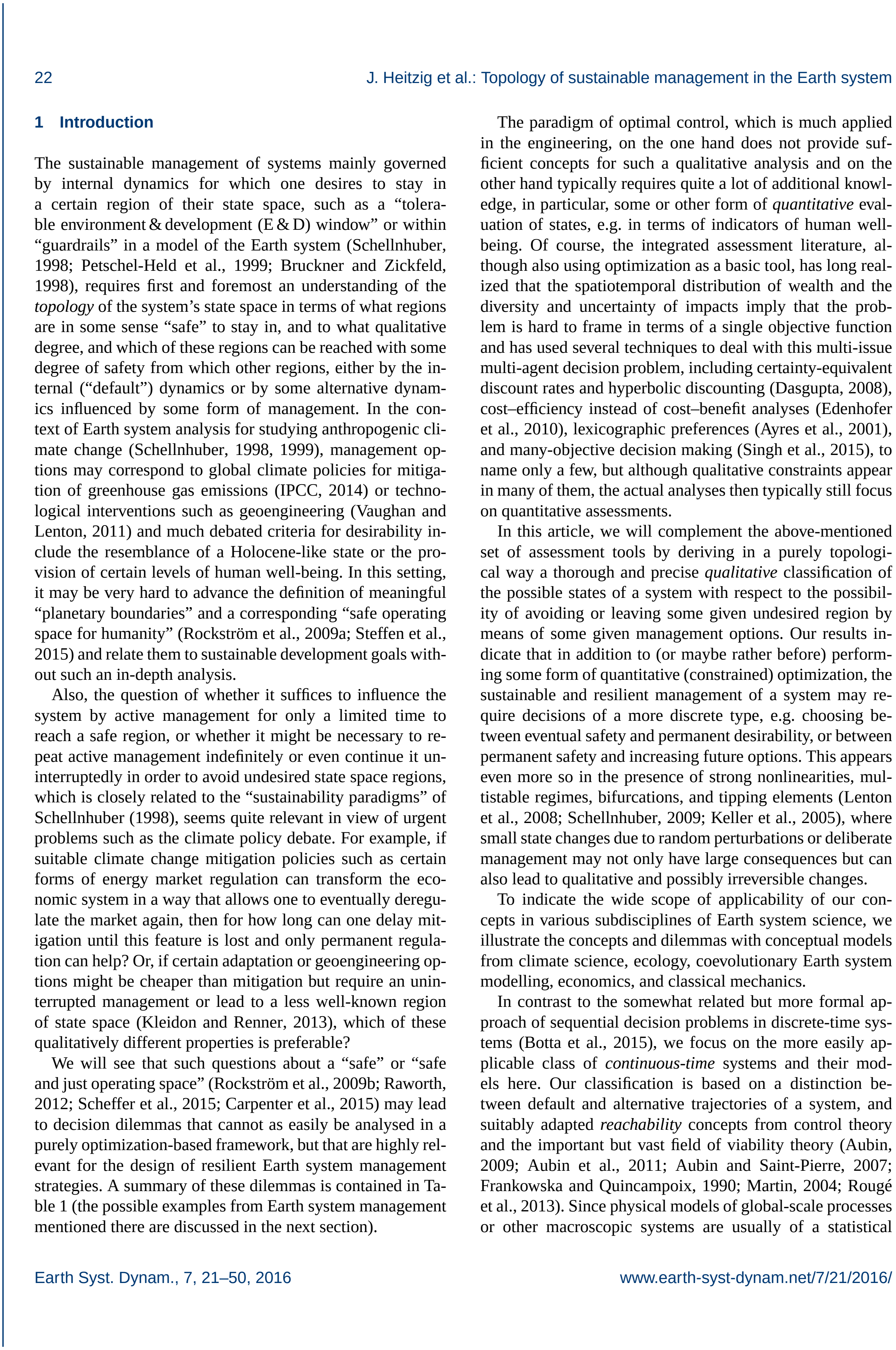
Abstract. To keep the Earth system in a desirable region of its state space, such as defined by the recently suggested "tolerable environment and development window", "guardrails", "planetary boundaries", or "safe (and just) operating space for humanity", one needs to understand not only the quantitative internal dynamics of the system and the available options for influencing it (management) but also the structure of the system's state space with regard to certain qualitative differences. Important questions are, which state space regions can be reached from which others with or without leaving the desirable region, which regions are in a variety of senses "safe" to stay in when management options might break away, and which qualitative decision problems may occur as a consequence of this topological structure?

In this article, we develop a mathematical theory of the qualitative topology of the state space of a dynamical system with management options and desirable states, as a complement to the existing literature on optimal control which is more focussed on quantitative optimization and is much applied in both the engineering and the integrated assessment literature. We suggest a certain terminology for the various resulting regions of the state space and perform a detailed formal classification of the possible states with respect to the possibility of avoiding or leaving the undesired region. Our results indicate that, before performing some form of quantitative optimization such as of indicators of human well-being for achieving certain sustainable development goals, a sustainable and resilient management of the Earth system may require decisions of a more discrete type that come in the form of several dilemmas, e.g. choosing between eventual safety and uninterrupted desirability, or between uninterrupted safety and larger flexibility.

We illustrate the concepts and dilemmas drawing on conceptual models from climate science, ecology, coevolutionary Earth system modelling, economics, and classical mechanics, and discuss their potential relevance for the climate and sustainability debate, in particular suggesting several levels of planetary boundaries of qualitatively increasing safety.

Published by Copernicus Publications on behalf of the European Geosciences Union.

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).



1 Introduction

The sustainable management of systems mainly governed by internal dynamics for which one desires to stay in a certain region of their state space, such as a “tolerable environment & development (E & D) window” or within “guardrails” in a model of the Earth system (Schellnhuber, 1998; Petschel-Held et al., 1999; Bruckner and Zickfeld, 1998), requires first and foremost an understanding of the *topology* of the system’s state space in terms of what regions are in some sense “safe” to stay in, and to what qualitative degree, and which of these regions can be reached with some degree of safety from which other regions, either by the internal (“default”) dynamics or by some alternative dynamics influenced by some form of management. In the context of Earth system analysis for studying anthropogenic climate change (Schellnhuber, 1998, 1999), management options may correspond to global climate policies for mitigation of greenhouse gas emissions (IPCC, 2014) or technological interventions such as geoengineering (Vaughan and Lenton, 2011) and much debated criteria for desirability include the resemblance of a Holocene-like state or the provision of certain levels of human well-being. In this setting, it may be very hard to advance the definition of meaningful “planetary boundaries” and a corresponding “safe operating space for humanity” (Rockström et al., 2009a; Steffen et al., 2015) and relate them to sustainable development goals without such an in-depth analysis.

Also, the question of whether it suffices to influence the system by active management for only a limited time to reach a safe region, or whether it might be necessary to repeat active management indefinitely or even continue it uninterruptedly in order to avoid undesired state space regions, which is closely related to the “sustainability paradigms” of Schellnhuber (1998), seems quite relevant in view of urgent problems such as the climate policy debate. For example, if suitable climate change mitigation policies such as certain forms of energy market regulation can transform the economic system in a way that allows one to eventually deregulate the market again, then for how long can one delay mitigation until this feature is lost and only permanent regulation can help? Or, if certain adaptation or geoengineering options might be cheaper than mitigation but require an uninterrupted management or lead to a less well-known region of state space (Kleidon and Renner, 2013), which of these qualitatively different properties is preferable?

We will see that such questions about a “safe” or “safe and just operating space” (Rockström et al., 2009b; Raworth, 2012; Scheffer et al., 2015; Carpenter et al., 2015) may lead to decision dilemmas that cannot as easily be analysed in a purely optimization-based framework, but that are highly relevant for the design of resilient Earth system management strategies. A summary of these dilemmas is contained in Table 1 (the possible examples from Earth system management mentioned there are discussed in the next section).

The paradigm of optimal control, which is much applied in the engineering, on the one hand does not provide sufficient concepts for such a qualitative analysis and on the other hand typically requires quite a lot of additional knowledge, in particular, some or other form of *quantitative* evaluation of states, e.g. in terms of indicators of human well-being. Of course, the integrated assessment literature, although also using optimization as a basic tool, has long realized that the spatiotemporal distribution of wealth and the diversity and uncertainty of impacts imply that the problem is hard to frame in terms of a single objective function and has used several techniques to deal with this multi-issue multi-agent decision problem, including certainty-equivalent discount rates and hyperbolic discounting (Dasgupta, 2008), cost–efficiency instead of cost–benefit analyses (Edenhofer et al., 2010), lexicographic preferences (Ayres et al., 2001), and many-objective decision making (Singh et al., 2015), to name only a few, but although qualitative constraints appear in many of them, the actual analyses then typically still focus on quantitative assessments.

In this article, we will complement the above-mentioned set of assessment tools by deriving in a purely topological way a thorough and precise *qualitative* classification of the possible states of a system with respect to the possibility of avoiding or leaving some given undesired region by means of some given management options. Our results indicate that in addition to (or maybe rather before) performing some form of quantitative (constrained) optimization, the sustainable and resilient management of a system may require decisions of a more discrete type, e.g. choosing between eventual safety and permanent desirability, or between permanent safety and increasing future options. This appears even more so in the presence of strong nonlinearities, multistable regimes, bifurcations, and tipping elements (Lenton et al., 2008; Schellnhuber, 2009; Keller et al., 2005), where small state changes due to random perturbations or deliberate management may not only have large consequences but can also lead to qualitative and possibly irreversible changes.

To indicate the wide scope of applicability of our concepts in various subdisciplines of Earth system science, we illustrate the concepts and dilemmas with conceptual models from climate science, ecology, coevolutionary Earth system modelling, economics, and classical mechanics.

In contrast to the somewhat related but more formal approach of sequential decision problems in discrete-time systems (Botta et al., 2015), we focus on the more easily applicable class of *continuous-time* systems and their models here. Our classification is based on a distinction between default and alternative trajectories of a system, and suitably adapted *reachability* concepts from control theory and the important but vast field of viability theory (Aubin, 2009; Aubin et al., 2011; Aubin and Saint-Pierre, 2007; Frankowska and Quincampoix, 1990; Martin, 2004; Rougé et al., 2013). Since physical models of global-scale processes or other macroscopic systems are usually of a statistical

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

Table 1. Preview of dilemma types discussed in the article.

Name	Option 1	Option 2	Possible example
“Glade” dilemma	higher desirability/flexibility	safety	adaptation/mitigation
“Lake” dilemma	uninterrupted desirability	eventual safety	great transformation
“Port” dilemma	higher flexibility	higher desirability	land-use change
“Harbour” dilemma	uninterrupted desirability	eventually higher desirability/flexibility	space colonization
“Dock” dilemma	uninterrupted safety	eventually higher desirability/flexibility	new technologies

physics nature in the sense that they represent the aggregate effects of many micro-scale processes by suitable approximations, their proper interpretation typically requires one to expect small (actually or seemingly) random perturbations. We take this into account here by strengthening the usual notion of reachability to one of *stable reachability*, and by requiring the featured subsets of state space to be topologically open (instead of closed) sets, so that infinitesimal perturbations cannot kick the system out of them.

In the next subsection (“Metaphorical framework”), we will briefly summarize our main concepts with the help of a metaphorical illustration, before introducing the corresponding formal notation in Sect. 2 in a concise way, reserving a more detailed formal treatment for Appendix A. The framework is then exemplified at the hand of several low-dimensional, conceptual models from various subdisciplines of Earth system science including climate science, ecology, and coevolutionary social–environmental Earth system modelling (Sect. 3) in order to indicate the wide scope of applicability of our concepts. A thorough analysis of more realistic and thus higher-dimensional models of the Earth system is something we have to leave for future studies since that would require further improvement of the numerical methods and algorithms employed for finding region boundaries. We conclude with a discussion and outlook in Sect. 4.

1.1 Metaphorical framework

As a start, let us take the common metaphor that “we’re all in the same boat” literally and represent the state of the Earth system with all its natural and socio-economic parts at each point in time by a single small boat floating or being rowed somewhere on a rather complex system of waters such as in Fig. 1.

The boat can only be on water, not on land, and will generally float along with the stream that represents the inherent dynamics of the Earth system over hundreds and thousands of years (the “default trajectory”), but it may also be rowed in more or less different directions depending on how strong the current of the stream is, and this possibility of rowing represents humankind’s agency in deliberately influencing the Earth system’s course to some extent by some or other form of what we will call “management” below. Let us assume that the main qualitative distinction with regard to where humanity wants their boat to be is represented by a division of

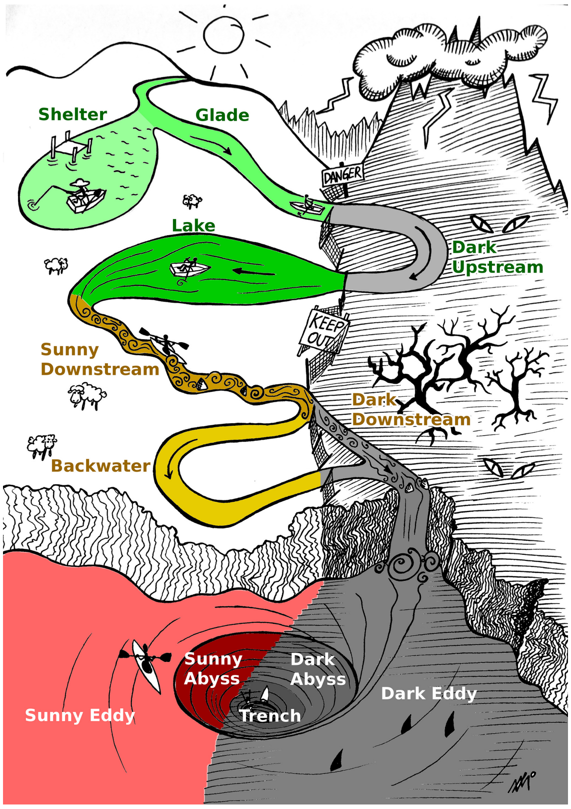


Figure 1. Metaphorical summary of concepts introduced in Sect. 1.1 (“Metaphorical framework”) inspired by Schellnhuber (1998). It depicts a river flowing from the mountains to the sea while going through sunny (left) and dark parts (right) where humanity can float and row on a boat. In the *shelter*, no rowing is needed to remain in the sun. One can row against the stream direction in slowly flowing parts, shown with long thin arrows, but in fast parts marked with swirls this is not possible. This setting gives rise to a number of qualitatively different regions of the system’s state space that can be found in any manageable dynamical system as well: *upstream* regions such as *glades* and *lakes* from where the shelter can be reached, *downstream* regions such as the *backwaters* from where one can at best stay in the sun by management, and several types of worse regions, all labelled here and explained in the text. See also Figs. 2 and 3.

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

the whole region into a desirable, “sunny” region on the left and an undesirable, “dark” region on the right, both containing several parts of the waters that may be connected in any imaginable ways, and with the natural water flow possibly drawing the boat back and forth between these two regions. The sunny region is meant to consist of all those possible states of the natural and socio-economic parts of the Earth system in which some generally agreed environmental and living standards are met, such as those defined by the human rights charter or the sustainable development goals (global goals) recently adopted by the United Nations. An alternative definition of the sunny region has been put forward in the planetary boundary framework (Rockström et al., 2009a; Steffen et al., 2015), where states lying within the corridor of Earth system variability during the Holocene that human societies are adapted to are considered as desirable.

We will show in this article that in such a setting, no matter how the waters look exactly, the general situation is in a certain sense always equivalent to the situation depicted in Fig. 1. There will in general be a certain sunny water region where one does not need to row at all in order to stay in the sun forever but can simply lean back and let the boat float around inside that region. In the picture, this region is the top-left tranquil tarn, but in general this region may also consist of several disconnected parts which we will call the *shelters* to emphasize their desirable and safe nature. Indeed, we will argue below that these shelters may be the most natural candidates for being called a “safe and just operating space for humanity”, only that we may not yet be in them. In the Earth system, there may be several such shelters, one of which might correspond to resilient states of the world (Folke et al., 2010) where humanity lives reconnected to the biosphere (Folke et al., 2011) and no active intervention or constant large-scale management is needed.

Connected to the shelter(s), there will in general also be other parts of the sunny region where it would not be safe to just lean back since the flow would then draw the boat into the dark after some time, but from where the shelters can still be reached by some suitable rowing, as show to the left of the “danger” sign in the image. For their “almost-safe” character, we will call such regions *glades*. If the glade is for some reason more desirable or offers more flexibility in terms of where one may row, one may face a *dilemma* when in a glade, i.e. a qualitative decision problem, namely whether to prefer staying in the safety of the shelter or in the more desirable but unsafe glade.

The shelters may also be reached by rowing from some places within the dark region (e.g. to the right of the “danger” sign) or through such a dark region from some other sunny places (such as those above the “keep out” sign). Among these latter sunny places from where the shelters can be reached only through the dark, there will generally be some places where one may alternatively stay forever in the sun by continuous rowing instead of passing through the dark and leaning back eventually. Such special places as the one above the “keep out” sign will be called *lakes* here, and they are characterized by a moderate current towards a dark place that one can row against and by the decision dilemma that results from the question of whether one should indeed do so or rather row to a shelter through the dark.

All these regions together will be called the *upstream* region for reasons that should become clear soon. In any system’s state space, the upstream consists of all states from which the shelters can be reached by management, and it is partitioned into one or several shelters, glades, dark upstream parts, lakes, and some remaining sunny upstream parts where it is not possible to stay in the sun forever. In Fig. 1, the upstream ends where the *rapids* left of the “keep out” sign begin since there the stream becomes so strong that it becomes impossible to row against it in order to eventually reach a shelter. Once the boat has left the upstream via such a rapid, there is no hope of leaning back eventually and staying in the sun, and for this reason the borders of the upstream may be called the “no-regrets planetary boundaries”, forming a middle level of a hierarchy of planetary boundaries we will suggest in Sect. 4.

Further down the stream there will typically be places where it is still possible to stay in the sun forever, only that one has to row over and over again to do so, such as in the slow-moving side branch below the “keep out” sign in the picture. Such regions, called *backwaters* here, are similar to lakes, only without the option of rowing to a shelter, so that the lake dilemma does not occur since the only chance one has is to row against the slow current to stay in the backwater. While the upstream was defined by being able to reach a shelter, the *downstream* is now defined as all places from where a backwater but not a shelter can be reached, including the backwaters, some dark parts such as the slow-moving dark part just right of the backwater in the picture, and maybe some remaining sunny downstream parts from where one may reach a backwater only through the dark. An example of a backwater could be a “machine world” where humanity can fully control nature to its very minute detail. While they can stay within the sunny region for infinite time through this management, there is no way of reaching a shelter anymore because the ecosystem has been changed irreversibly.

The waterfall in Fig. 1 indicates that besides the upstream and downstream regions, where it is possible to stay in the sun eventually, there will in general be further, less hopeful places the system may be in, from where one cannot avoid entering the dark over and over again. In some of those, one can at least make sure that one also spends some time in the sun over and over again, as depicted by the kayak in the picture. Since this is typically connected to some form of cyclic motion, we will call such regions *eddies*. In some eddies, failing to row correctly may push the boat into an even less desirable region, called an *abyss*, from where one can no longer avoid ending up in the dark forever eventually, as in the ring-shaped abyss shown inside the eddy in the figure. Finally, the dark region from where there is no escape, depicted in the centre of the abyss, will be called a *trench*.

This completes our main partitioning of the Earth system's or any other manageable system's state space into qualitatively different regions: upstream and downstream, defined by being able to reach shelters or backwaters; abysses, defined by not being able to avoid ending up in a trench; and eddies in between, defined by being at least able to switch between sun and dark forever. Figure 2 summarizes all these regions in the form of a decision tree, where one can identify the region the system is in by answering a small number of questions. That our partitioning is indeed complete and can be given a suitable and unambiguous mathematical form for all kinds of systems is shown in the next section.

While in Fig. 1, each of the introduced set of system states is just one topologically connected region, in general most of these sets are composed of several disjoint regions, so there may be several shelters, glades, lakes, etc. On a finer level, these may be analysed further by looking at which parts may be reached from which other parts, and this leads to a finer, hierarchical partition into *ports*, *rapids*, *harbours*, *docks*, etc. and to several new types of dilemmas, as shown in Fig. 3.

All of the five types of dilemmas listed in Table 1 can easily occur in the collective "management" or governance of the Earth system by humanity. A glade dilemma may occur if adaptation is seen as preferable to mitigation for welfare reasons but turns out to be a riskier option due to a higher uncertainty of the corresponding climate impacts. A lake dilemma can arise if a great transformation of the global energy system towards a carbon-free economy would temporarily lead to welfare losses in poorer countries. A port dilemma may come from the option of increasing welfare by extending industrial agriculture causing biodiversity loss (decreasing flexibility) due to the related large-scale land-use change. A harbour dilemma could occur in the future when colonization of other planets (increasing flexibility) becomes feasible but extremely costly. Finally, a dock dilemma arises whenever a very promising new technology with some unknown risks and side effects (such as genetically engineered food production) could be introduced on a planetary scale.

2 Formal framework

We will now put all of the above on thorough mathematical footing. Let us assume a *manageable dynamical system with desirable states*, given by the following components:

i. a dynamical system with a *state space* X, *default dynamics* represented by a family of *default trajectories* $\tau_x(t)$, and some basic *topology* on X (e.g. the Euclidean topology; see Appendix A1 for more detail);

ii. a notion of *desirable states* represented by an open set $X^+ \subseteq X$, called the *sunny region*, whose complement $X^- = X - X^+$ we call the *dark*;

iii. a notion of *management options* represented by a family $\mathcal{M}_x$ of *admissible trajectories* μ for each $x \in X$.

We assume that one can switch immediately to any trajectory $\mu \in \mathcal{M}_x$ whenever in state x. We say the system *floats* when it follows a default trajectory, and that we may *row* the system along any other admissible trajectory.

Note that although, formally, we consider deterministic autonomous systems only, non-deterministic systems can be incorporated by considering probability distributions as states, time-delay systems can be treated similarly, and externally driven or otherwise explicitly time-dependent systems can be covered by including time t as a variable with $\dot{t} = 1$ into the state vector. Also, if management involves some form of inertia, e.g. if not the propelling vector $\boldsymbol{v}$ of a boat but only its acceleration $\dot{v}$ can be changed discontinuously, the proper way to model this in our framework would be to treat $\boldsymbol{v}$ as part of the state.

2.1 Qualitative distinction of regions with regard to sustainable manageability of desirability

The main idea of the coarsest of our classifications of states is to first identify (i) a *safe* region where management is unnecessary, called the shelters S, and (ii) a less safe but larger manageable region M where one can permanently avoid the dark at least by management. Then we classify all states with regard to whether and how X^+, S, and M can be stably reached from the current state by management. For each state, we ask the following questions. (iii) Can S be stably reached, and if so, can the dark be avoided on the way? (iv) If not, can M be stably reached? (v) If not, can we stably reach X^+ over and over again, or at least once again? We will see that these criteria lead to a partition of state space into a "cascade" consisting of five main regions: upstream U, downstream D, eddies E, abysses Υ, and trenches Θ. Each of these will then be split up further into sets such as glades G, lakes L, and backwaters W by asking further qualitative questions. In choosing these figurative terms, we try to avoid a too technically sounding language and rather extend the useful and common metaphor of "flows" and "basins" in a natural way without trying to match their common-language meanings too accurately.

To acknowledge the fact that all real-world dynamics and management will be subject to at least infinitesimal noise and errors, we base the formal definition of these state space regions on certain notions of *invariant open kernel, sustainability*, and *stable reachability*, whose symbolic mathematical definitions and algebraic properties are detailed in Appendix A2.

2.2 Shelters, manageable region, upstream, and downstream

The *invariant open kernel* of a set $A \subseteq X$, denoted $A^{\iota\circ}$, is the largest open subset of A that contains the default trajecto-

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

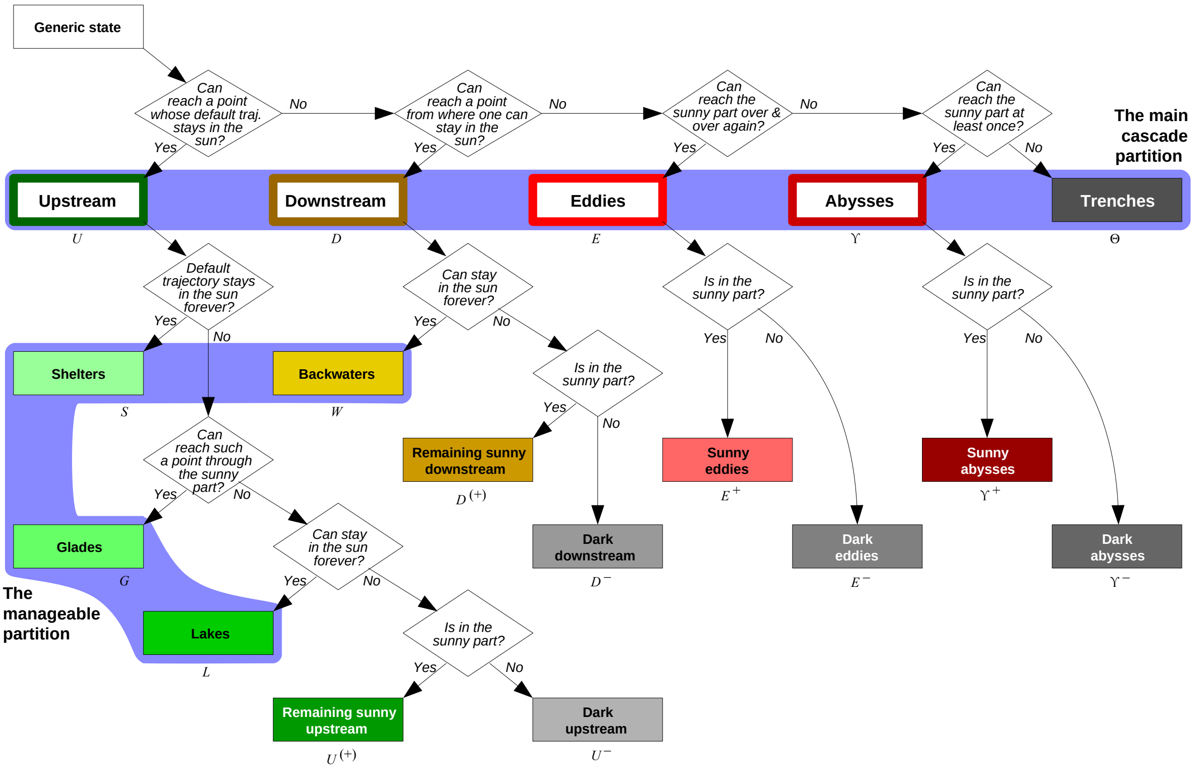


Figure 2. Decision tree summarizing the partition of a manageable dynamical system's state space with regard to stable reachability of the desired region or the shelters (main cascade), and the finer partition of the manageable region. The colour scheme (grey undesired regions, green upstream regions, yellow downstream regions, red eddies, and abysses, with lighter meaning better) is also used in the remaining figures.

ries of all its own points. The *shelters* are the invariant open kernel of the sunny region,

$$S = \left(X^+\right)^{\circlearrowleft}. \tag{1}$$

S contains all sunny states whose default trajectories stay in the sunny region X^+ forever without any management even when infinitesimal (or small enough) perturbations occur. In other words, when inside S, one *will* "stably" stay in X^+ *by default*.

We call an open set A *sustainable* (in the basic sense of the word, simply meaning that it can be sustained) iff it contains an admissible trajectory for each of its points. The *sustainable kernel* of a set $A \subseteq X$, denoted $A^{\mathcal{S}}$, is the largest sustainable open subset of A. We call the sustainable kernel of the sunny region the *manageable region*:

$$M = \left(X^+\right)^{\mathcal{S}} \supseteq S. \tag{2}$$

In other words, when inside M, one *can* stably stay in X^+ *by management*.

In Appendix A2, we introduce a suitable notion of stable reachability to overcome two problems with the classical notion of (plain) reachability known from control theory. For now, let us assume we know what we mean when saying that a state y or a set $Y \subseteq X$ is *stably reachable* from some state x *through* some set $A \subseteq X$, denoted $x \rightsquigarrow_A y$ or $x \rightsquigarrow_A Y$. Using this notion of stable reachability for the choice $A = X$ (other choices of A will be used in the next section), we can now define the upstream U as the set of states from where the shelters S can be stably reached at all. Likewise, the downstream D consists of all states from which the manageable region M but not the shelters can be stably reached:

$$U = (\rightsquigarrow_X S) \supseteq S, \tag{3}$$

$$D = (\rightsquigarrow_X M) - (\rightsquigarrow_X S) = (\rightsquigarrow_X M) - U \supseteq M - U. \tag{4}$$

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

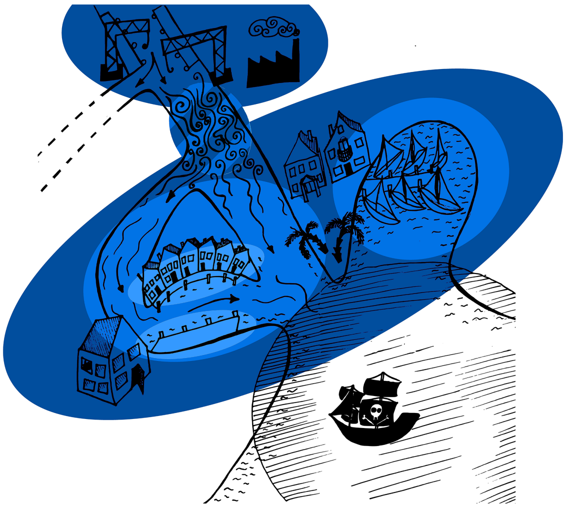


Figure 3. Illustration of port, harbour, and dock dilemmas introduced in Sect. 1.1 ("Metaphorical framework"). As in Fig. 1, humanity can float in and row a boat on a complex waterway. From the upper *port* city (upper dark-blue region), one can get to some unknown region to the left and to another, nicer port city (lower dark blue) at the shore through a rapid (hatched blue) which cannot be traversed in the other direction. This choice between desirability and flexibility forms a *port dilemma*. The nicer port city has two harbours (middle blue regions), of which the right one is more desirable, and between which one can switch only through an undesired region where pirates loom (circular area). Boats in the left harbour face the *harbour dilemma* of choosing between either avoiding the undesired region by all means or eventually reaching a place of higher desirability. Finally, in the left harbour there are two safe *docks* (light-blue regions), of which the top one is more desirable, and between which one can switch only through an unsafe part of the harbour from which one may be drawn into the undesired region if the engine fails. Boats in the bottom dock face the *dock dilemma* of choosing between uninterrupted safety and eventual higher desirability.

2.3 Trenches, abysses, eddies, and the main cascade

On the other, dark end of what we will call the main cascade, we first define the trenches Θ as that region in the dark from which one cannot stably reach the sunny region even once,

$$\Theta = X - \left(\rightsquigarrow_X X^+\right) \tag{5}$$

(this concept approximately corresponds to the "catastrophe domains" of Schellnhuber, 1998).

Now we turn to the region from where one cannot avoid ending up in the trenches. We define the abysses Υ as the closure of this region, minus the trenches:

$$\Upsilon = \overline{\{x \in X | \forall \mu \in \mathcal{M}_x \exists t \geqslant 0 : \mu(t) \in \Theta\}} - \Theta. \tag{6}$$

The closure is taken since even an infinitesimally small perturbation from a point in this closure can make the trenches unavoidable.

Finally, the eddies E are the remainder of X, i.e. the part from where the manageable region cannot be stably reached but the trenches can be avoided:

$$\begin{aligned} E &= X - U - D - \Upsilon - \Theta \\ &= (X - (\rightsquigarrow_X M)) \cap (X - (\Upsilon + \Theta)). \end{aligned} \tag{7}$$

Thus, when in the eddies, even though one can reach the sunny part over and over again, one cannot stay there forever but has to visit the dark repeatedly.

A connected component of Θ, Υ, or E will be called an individual trench, abyss, or eddy, and the latter two typically have sunny and dark parts.

The system $\mathcal{C} = \{U, D, E, \Upsilon, \Theta\}$ is a partition of X which we call the *main cascade* because of the following mutual reachability restrictions:

$$\neg(\Theta \rightsquigarrow \Upsilon), \neg(\Upsilon \rightsquigarrow E), \neg(E \rightsquigarrow D), \neg(D \rightsquigarrow U). \tag{8}$$

In other words, one might at best be able to go in the "downstream" direction by default or by management, from upstream to downstream to the eddies to the abysses to the trenches, but not in the other, "upstream" direction (see also Fig. 2).

2.4 The glades and lake dilemmas, backwaters, and the manageable partition

Some of the states in the manageable region M may be in $U = (\rightsquigarrow_X S)$ but not in $(\rightsquigarrow_{X+} S)$. This motivates the definition of two subsets of M via the relation of sunny stable reachability, $\rightsquigarrow_{X+}$, namely (i) the glades G, from where the shelters can be stably reached through the sun, and (ii) the lakes L, from where the shelters can be stably reached only through the dark:

$$G = (\rightsquigarrow_{X+} S) - S, \tag{9}$$

$$L = M \cap U - (\rightsquigarrow_{X+} S) = M \cap U - S - G. \tag{10}$$

Glades and lakes are two particularly interesting types of regions since in both one has a qualitative decision problem. The *glade dilemma* occurs if a glade is for some reason more desirable than its shelter, since then one has to decide whether to stay in the more desirable but unsafe glade or row to the less desirable but safe shelter. The *lake dilemma* exists in every lake: shall one stay in the sun by rowing over and over again, but risking to float into the dark if the paddle breaks, or shall one move into a shelter, accepting a temporary passage through the dark, to be able to recline in safety eventually? In other words, the lake dilemma is a choice between uninterrupted desirability and eventual safety. Below we will encounter more qualitative dilemmas of this and other types.

While $\{S, G, L\}$ is a partition of $M \cap U$, the downstream D may also contain a manageable part, the backwaters W. This is the region where one may stay in the sun forever by

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

rowing over and over again, but where one may not stably reach the shelters at all, not even through the dark:

$$W = M \cap D = M - U. \tag{11}$$

This completes the *manageable partition*

$$M = S + G + L + W. \tag{12}$$

Also, both U and D may contain points outside M, which we call the *dark upstream/downstream*,

$$U^- = U \cap X^-, \quad D^- = D \cap X^-, \tag{13}$$

and the *remaining sunny upstream/downstream*,

$$U^{(+)} = \left(U \cap X^+\right) - M, \quad D^{(+)} = \left(D \cap X^+\right) - M, \tag{14}$$

leading to the *upstream* and *downstream* partitions

$$\begin{aligned} U &= S + G + L + U^{(+)} + U^-, \\ D &= W + D^{(+)} + D^-. \end{aligned} \tag{15}$$

Finally, one can divide the eddies and abysses into sunny and dark parts:

$$E^\pm = E \cap X^\pm, \quad \Upsilon^\pm = \Upsilon \cap X^\pm. \tag{16}$$

All the sets introduced so far are summarized in Fig. 2 in the form of a decision tree that allows for a fast classification of individual states.

2.5 Finer distinction of regions with regard to mutual reachability of different types

In addition to the glade and lake dilemmas introduced above, there exist at least three further types of qualitative decision problems, all related to the question of which parts or subregions of the above introduced regions may be stably reached from which other parts, and whether corresponding transition pathways exist that do not leave the shelters or at least the sunny region, or only through the dark. In order to study these questions, we introduce three additional, successively finer partitions derived from the reachability relations $\rightsquigarrow_X$ (stable reachability) and $\rightsquigarrow_{X^+}$ (stable reachability through the sun) that we used already above, and from the even more restrictive relation $\rightsquigarrow_S$ (stable reachability through the shelters).

2.5.1 The ports-and-rapids partition and network, and the port dilemma

While from each state in U, one can stably reach some part of S, one cannot in general navigate freely inside S or U or any other member of the main cascade $\mathcal{C}$. Let us call a maximal region in which one can navigate freely a *port* (see Appendix A3 for more thorough formal definitions and proofs of the claimed properties). Each port is completely contained in one of the sets U, D, E, Υ^-, Θ, and none can intersect Υ^+, so the notion of ports fits well into the hierarchy of regions that began with the main cascade and the manageable partition. But there are also *transitional* states not belonging to any port since one cannot return to them. Thus, to extend the system of all ports into a partition of all of X, we also have to classify these non-port states, and we do so by asking which ports they can reach and from which ports they can be reached. States that are equivalent in this sense form what we call a *rapid*. It turns out that U and D are then partitioned into ports and rapids, and so is each individual eddy, abyss, and trench. The reachability relations between ports and rapids form a directed network that concisely summarizes the overall structure of all management options.

Figure 1 shows the very simple case of a linear network: the whole upstream is one port, the sunny downstream and the adjacent fast-moving part of the dark downstream form a rapid, the backwater and the slow-moving part of the dark downstream form another port, the waterfall is another rapid, the eddy is a port again, and the abyss and the trench are rapids. In the examples below, we will, however, see that much more complex ports-and-rapids networks may occur in models, and one can prove that any acyclic graph may occur as the ports-and-rapids network of some system.

The ports-and-rapids partition is helpful in the discussion of a certain type of dilemma that results from two different objectives which may not be easily balanced: (i) the objective of being in or reaching a state with high *intrinsic desirability*, e.g. as measured by some qualitative preference relation finer than the mere distinction between "desirable" and "undesirable", or even by some quantitative evaluation such as a welfare function, and (ii) the objective of retaining an amount of *flexibility* as large as possible by being in or reaching a state from which a large part of state space is reachable. Flexibility may be important in particular in situations in which there is some uncertainty about future management options and/or future preferences (Kreps, 1979). We call this a *port dilemma*.

2.5.2 The harbours-and-channels partition and network, and the harbour dilemma

Since they do not take into account the definition of the desirable region X^+ at all, ports and rapids are not directly compatible with the regions from the manageable partition $\mathcal{M}$ since their members may overlap in complex ways. However, we can construct a very similar but finer partition based on stable reachability through the sun ($\rightsquigarrow_{X^+}$) instead of (plain) stable reachability, restricted to the sunny region, and the result turns out to be compatible with $\mathcal{M}$.

A maximal region in which one can freely navigate without leaving the sun is called a *harbour*. A region of states that do not belong to any harbour but from which the same harbours can be reached through the sun and which can be reached from the same harbours through the sun is called a *channel*. Since each harbour or channel lies completely in a port or a rapid, the harbours and channels form a finer partition than the ports and rapids and form a finer layer of the reachability network in which the links represent reachability through the sun instead of mere reachability.

The harbours-and-channels partition allows one to identify decision problems involving (i) the objective of *staying* in a desirable state and (ii) the objective of eventually *reaching* a state with higher desirability or flexibility, which is called a *harbour dilemma* here.

2.5.3 The docks-and-fairways partition and network, and the dock dilemma

Note that although the harbours-and-channels partition is finer than that into ports and rapids, there is still one important region that can have nontrivial overlaps with harbours and channels, namely the shelters S. In order to complete our hierarchy of partitions and networks of regions, we therefore introduce a third and finest partition and network level, restricted to S, based on the notion of *stable reachability through the shelters*, $\rightsquigarrow_S$.

In complete analogy to the above, a maximal region of states that are mutually reachable through S is called a *dock*, and the non-dock states in S are classified into so-called *fairways* with regard to their reachability of these docks. Again, each dock or fairway lies completely in a harbour or channel, and they form a third layer of the reachability network whose links now represent the safest form of reachability, namely through the shelters.

Finally, the docks-and-fairways partition is helpful in the discussion of dilemmas involving (i) the objective of staying in a *safe* state (i.e. in the shelters) and (ii) the objective of eventually reaching a state with higher desirability or flexibility. We call this a *dock dilemma*.

2.6 Summary of the introduced hierarchy of partitions and networks

To summarize, we have now a hierarchy of ever-finer partitions of the system's state space at our hands. We began with the main cascade $\mathcal{C} = \{U, D, E, \Upsilon, \Theta\}$, its refinement into the partition $\{S, G, L, U^{(+)}, U^{-}, W, D^{(+)}, D^{-}, E^{+}, E^{-}, \Upsilon^{+}, \Upsilon^{-}, \Theta\}$ (see Fig. 2), and the further refinement by topological connectedness into individual shelters, glades, lakes, backwaters, eddies, abysses, and trenches. These partitions represent the qualitative differences in stable reachability of the shelters or the manageable set, thus allowing for a first classification of states with regard to the possibilities of sustainable management, and may reveal decision problems of the type of glade or lake dilemma which will occur in many of the examples below, where one has to choose between higher safety and higher desirability or flexibility or between uninterrupted desirability and eventual safety.

A different refinement of $\mathcal{C}$ into the ports-and-rapids network is still based on stable reachability alone but contains other details suitable for the identification and discussion of possible port dilemmas that involve a choice between higher desirability and higher flexibility. Inside the desirable region X^{+}, this partition can be refined into the harbours-and-channels network suitable for the discussion of harbour dilemmas that involve a choice between uninterrupted desirability and eventually higher desirability or flexibility, and further into the docks-and-fairways network suitable for the discussion of dock dilemmas that involve a choice between uninterrupted safety and eventually higher desirability or flexibility (Table 1).

These three networks may also be interpreted as a three-level "network of networks" with nodes representing state space regions of different quality and size. A network-theoretic analysis of it using methods such as the node-weighted measures of Heitzig et al. (2012) may especially be interesting in the context of varying system parameters and bifurcations such as those in Fig. B2, but this is beyond the scope of this article.

3 Examples

In this section, we will apply the introduced framework to several illustrative examples from natural and coevolutionary Earth system modelling, ecology, socio-economics, and classical mechanics. The examples have been chosen not for their realism but for their simplicity in order to show the broad scope of potential applicability of our concepts, as well as the relevance of the identified types of decision dilemmas in both the natural and socio-economic components of the Earth system.

3.1 Carbon cycle and planetary boundaries

Our first example is from natural Earth system modelling and illustrates which of the above-introduced regions occur most often for systems that possess only a single, globally stable, and desirable attractor.

Anderies et al. (2013) proposed a conceptual model of the global carbon cycle capturing its main features while keeping the model sufficiently low-dimensional to be able to discuss the planetary boundaries concept with it. We use their model for pre-industrial times, which has three dynamical variables c_{m}, c_{t} and $c_{\mathrm{a}} = 1 - c_{\mathrm{m}} - c_{\mathrm{t}}$ representing the maritime, terrestrial, and atmospheric shares of the fixed global carbon stock. The dynamics are of the form

$$\dot{c}_{\mathrm{m}} = a_{\mathrm{m}} (c_{\mathrm{a}} - \beta c_{\mathrm{m}}), \quad \dot{c}_{\mathrm{t}} = f(c_{\mathrm{a}}, c_{\mathrm{t}}) - \alpha c_{\mathrm{t}},$$

where a_{m} and β are diffusion parameters, f is a function representing photosynthesis and respiration, and α governs the human offtake rate from the terrestrial carbon stock. See Anderies et al. (2013) for details and parameter values.

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

Since the parameter α can be considered the natural human management option for this system, we assume the default flow has a value of $\alpha = \alpha_+ = 0.5$, while management can reduce it by half to $\alpha = \alpha_- = 0.25$, which results in the trajectories shown in Fig. 4. Both have a unique stable fixed point in the interior of the state space which is globally attractive for all states with $c_t > 0$.

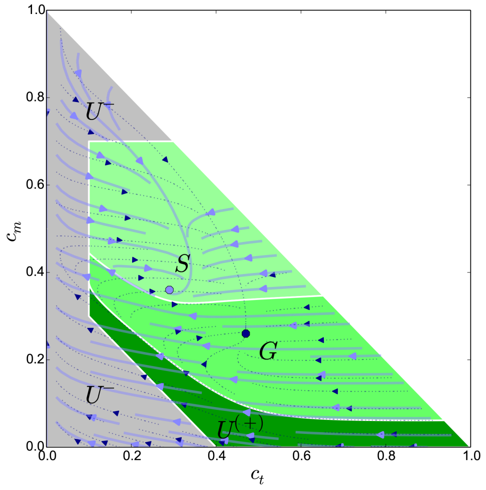


Figure 4. Phase portrait of the pre-industrial carbon cycle model of Anderies et al. (2013). Arrows indicate default/unmanaged dynamics (pale blue) and alternative/managed dynamics (dotted dark blue) from reducing the human offtake rate by half. Filled dots: corresponding stable fixed points. Grey area: undesired region defined by (i) upper bounds for maritime carbon c_m (white horizontal line, representing a planetary boundary related to ocean acidification) and atmospheric carbon $1 - c_t - c_m$ (white diagonal line, related to a climate change boundary) and a lower bound for terrestrial carbon c_t (white vertical line, representing an ecosystem services planetary boundary). Coloured areas and labels: derived state space partition (see text); colours as defined in Fig. 2: a shelter S around the globally stable fixed point of the default dynamics, a glade G from where S can be reached by management without violating the bounds, and a remaining sunny upstream $U^{(+)}$ from where one cannot avoid violating the bounds temporarily.

In order to roughly represent the planetary boundaries relating to climate change, biosphere integrity, and ocean acidification (Rockström et al., 2009b; Steffen et al., 2015), we require a "sunny" state to have sufficiently low atmospheric carbon, at least a minimum value of terrestrial carbon, and not too large maritime carbon, leading to a dark region of the shape shown in Fig. 4 in grey. If, as shown, the unmanaged fixed point is sunny, one obtains a purely upstream situation with a shelter surrounding the fixed point, a glade, and a remaining sunny upstream $U^{(+)}$ as shown in the figure. For our (quite arbitrarily) chosen parameter values, a trajectory starting in the sunny upstream is likely to first cross the climate boundary and then the biosphere boundary before getting back into the sunny region, whereas it seems quite unlikely to cross the acidification boundary.

In this example, all non-upstream regions are empty, and so is the lake region; hence, no lake dilemma occurs. On the other hand, if one considers a higher c_t to be preferable, we get an example of the glade dilemma since the managed fixed point in the less safe glade has higher c_t than the unmanaged fixed point in the safer shelter. Note that this is neither a port, harbour, or dock dilemma since both points are in the same port and harbour and only the unmanaged one is in a dock.

If, instead, we had chosen the minimum value for c_t to be larger than the unmanaged equilibrium value, the shelter would be empty and the whole situation would change from upstream-only to either a downstream-only or an abyss-and-trench situation. This type of *topological bifurcation* will be studied in Sect. 3.4. In the next example, we will see a lake dilemma instead of a glade dilemma.

3.2 Competing plant types and multistability

The second example, from ecology, demonstrates how the lake dilemma may occur in a multistable system with a sunny and a dark attractor.

In this fictitious example, two plant types (1 and 2) compete for some fixed patch of land, modify the soil, and are harvested. Their growth follows logistic-type dynamics, with land cover proportions $x_{1,2} \in [0, 1]$ following the equations

$$\dot{x}_1 = x_1\left(K_1\left(x_{1,2}\right) - x_1\right) - h_1 x_1,$$
$$\dot{x}_2 = r x_2 (K_2(x_{1,2}) - x_2) - h_2 x_2.$$

In this, $r > 1$ is a constant productivity quotient, $h_{1,2}$ are the harvest rates, and the two dynamic capacities $K_1(x_{1,2}) = \sqrt{x_1}(1 - x_2) \leqslant 1$ and $K_2(x_{1,2}) = \sqrt{x_2}(1 - x_1) \leqslant 1$ represent the fact that each type modifies the soil quickly to its own benefit but to the other type's disadvantage (see Supplement 1 for a discussion of the model design based on Bever (2003), Kourtev et al. (2002), Kulmatiski et al. (2011), Levine et al. (2006), Poon (2011), and Read et al. (2003).

For our illustration, we assume that, on the default trajectories, both harvest rates $h_{1,2}$ equal some rather high value h_+, leading to low equilibrium harvests. We assume management can repeatedly choose between this default and two types of alternative trajectories. Type 1 has a lower value for both harvest rates, $h_{1,2} = h_- < h_+$, representing management by restricting harvests politically in order to yield higher long-term harvests, but without aiming to change the plant mix, as depicted in Fig. 5 (left panel). Type 2 management option has harvest rates $h_2 = 0$ and $h_1 = 2h_+$, representing management by temporarily protecting type 2 in order to change the plant mix to the higher productivity plant; we assume that this moratorium results in more intense harvesting of type 1, as depicted in Fig. 5 (right panel). We assume that both options exist simultaneously at all times (the separate plots of Fig. 5 are only for better discernibility of the trajectories). We

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

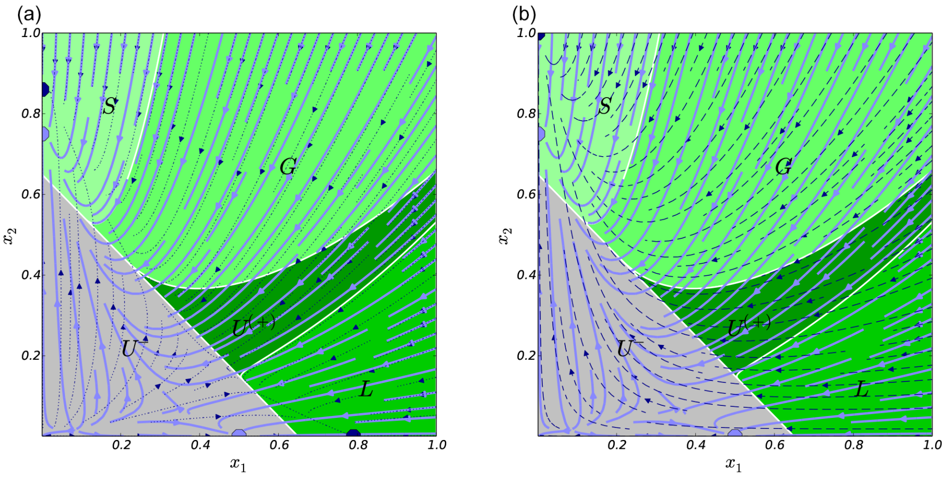


Figure 5. Competing plant types example, showing all upstream regions and illustrating the lake dilemma. A bistable system of two competing plant types with two simultaneous management options (depicted in separate plots only for discernibility). Management by a general harvesting quota (dotted arrows shown left) can ensure desirable long-term harvests of the less productive type x_1 (*lake* L). Management by temporary protection of the more productive type x_2 (dashed arrows shown right) can cause a transition to the desirable fixed point (in the *shelter* S), but only through the undesired region of low harvests (grey region). The state space partition boundaries resulting from both options together (white curves) and a desirable minimum harvest boundary (white diagonal) follow some admissible trajectory at each point.

set the desirable region to where $x_1 + x_2 > \ell$ for some $\ell > 0$ in order to ensure some minimum harvests.

For the choice $r = 2$, $h_+ = 0.2$, $h_- = 0.1$, $\ell = 0.65$ of the figure, the desirable high-productivity stable fixed point of the default dynamics at $\approx (0, 0.79)$ is in the sunny region and is thus contained in a shelter S. The latter is delimited by the default trajectory that meets the boundary to the undesired region tangentially. S can be stably reached from all states with $x_2 > 0$, and hence the upstream is $U = \{(x_1, x_2) | x_2 > 0\}$. The border of the glade G next to S can be found by backtracking the "widest" admissible trajectory that meets the boundary to the undesired region tangentially; this turns out to be a type 2 management trajectory as seen in Fig. 5 (right panel). This shows how the boundaries of regions may often be found by identifying tangential or otherwise significant points and backtracking the default and alternative trajectories leading to them.

The lower-productivity stable fixed point of the default dynamics (with $h_{1,2} = h_+$) at $\approx (0.52, 0)$ is undesired for this choice of X^+. From it one cannot only navigate to S but can also (and faster) get to the higher productivity stable fixed point of the first type of *managed* dynamics with $h_{1,2} = h_-$, at $\approx (0, 0.79)$, and stay there as long as management holds. Hence the region around $(0, 0.79)$ is part of the manageable region M. The exact boundary of this region (which soon turns out to be a lake, L) is the "widest" admissible trajectory that meets the boundary to the undesired region tangentially; in this case, this trajectory turns out to be a type 1 management trajectory as seen in Fig. 5 (left panel). To get from this type 1-dominated region to the type 2-dominated shelter S via the other management option of protecting type 2, one has to cross the undesired middle region in which both types coexist at a low level due to soil conditions that are suboptimal for both types. Hence the region around $(0, 0.79)$ is a lake. The associated lake dilemma is similar to a glade dilemma in that staying in a lake is unsafe as in a glade, but it differs in the reason why one may want to stay there: while staying in a glade may be attractive simply because the glade may be more desirable than the shelter in some quantitative sense, staying in a lake may seem attractive since that avoids having to pass through the dark to reach safety.

This form of the lake dilemma can also occur in other multistable systems when one of the attractors is in the dark but sufficiently close to the sunny region so that constant management can sustain the system in a sunny place near that attractor, and when other management options may push the system towards another, sunny attractor after crossing the dark.

Note that, in this example, the lake dilemma falls together with a port dilemma since after leaving the lake for the shelter, one cannot return. If we choose a slightly larger sunny region by lowering ℓ to $\ell = 0.45$, the unmanaged fixed point with $y = 0$ gets into X^+ and the former lake around it now becomes a second shelter, which might be called a *shelter–lake transition*. But from this shelter the other, more desirable shelter can still only be reached through the dark. Since the two shelters correspond to two harbours in the reachability network, this means the former lake dilemma has been converted into a harbour dilemma.

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

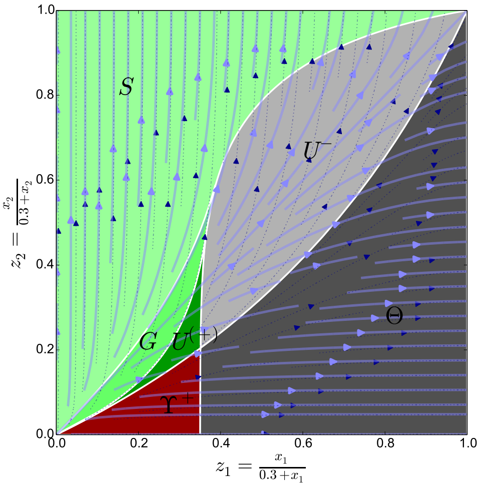


Figure 6. Substitution of a dirty technology. Coevolution of the cumulative production of a dirty technology (x_1) and a clean one (x_2) without (pale-blue curves) and with (dotted dark-blue curves) a subsidy for the clean technology. Undesired region with too high future usage of the dirty technology coloured in grey. Knowledge stocks $x_{1,2}$ were transformed to $z_{1,2} = x_{1,2}/(0.3 + x_{1,2})$ in order to capture their divergence to $+\infty$.

The example also shows that the more management options exist, the less trivial it is to find the boundaries between regions even in two-dimensional systems. For higher dimensions, one will usually have to rely on specialized numerical algorithms such as the viability kernel algorithm of Frankowska and Quincampoix (1990) from viability theory.

3.3 Substitution of a dirty technology

Our third example concerns a purely socio-economic part of the Earth system that bears some similarity to the preceding example but features regions from both ends of the main cascade: upstream and abyss/trench, without having the intermediate regions of downstream and eddies.

Instead of plants, in this example a certain produced good (e.g. electric energy) comes in two types which are economically perfectly substitutable but whose production processes use two different technologies – one "dirty" and one "clean" (e.g. conventional and renewable energy). The production costs C_1 and C_2 are convex functions of production output per time y_i and decrease over time via learning-by-doing dynamics that are similar to Wright's law (Nagy et al., 2013):

$$C_i\,(y_i) = \gamma_i y_i^{1+\sigma_i} / (1+\sigma_i) x_i^{\alpha_i}.$$

In this, x_i is cumulative past production (with $\dot{x}_i = y_i$), γ_i are cost factors, $\sigma_i > 0$ are convexity parameters, and $\alpha_i > 0$ are learning exponents. We assume that demand D depends linearly on price, $D(p) = D_0 - \delta p$, $\delta > 0$; that demand equals production, $D = y_1 + y_2$ ("market clearance"); and that price equals marginal costs, $p = \partial C/\partial y_i = \gamma_i\, y_i^{\sigma_i}/x_i^{\alpha_i}$, due to perfect competition among producers. One can then uniquely solve for the produced amounts y_i, getting some formula $y_i = f_i(x_1, x_2)$. This results in a two-dimensional dynamical system with state variables x_1, x_2 and equations

$$\dot{x}_i = f_i\,(x_1, x_2)\,.$$

Let us put $D_0 = 1$, $\delta = 1$, $\sigma_i \equiv 1/5$, $\alpha_i \equiv 1/2$, and assume that the default dynamics have $\gamma_i \equiv 1$, so that the long-term default behaviour is $p(t) \to 0$, $D(t) \to 1$. If the dirty technology (1) is the traditional one, so that $x_1(0) > x_2(0)$, we have $x_1(t) \to \infty$, $x_2(t) \to \hat{x}_2 < \infty$, $y_1(t) \to 1$, and $y_2(t) \to 0$, i.e. usage of the clean technology (2) will die out. If instead $x_1(0) < x_2(0)$, technology 1 will die out. Hence the system is bistable as in the plant example, but with attractors at infinity. To depict the diverging behaviour, we used the transformation $z_i = x_i/(0.3 + x_i)$ in Fig. 6.

The main dynamical difference to the plant example is, however, not the diverging behaviour, but has to do with the choice of management options. While in the plant example, the choice of management options led to an upstream-only situation in which the more desirable fixed point could be reached from everywhere, in this example we will get regions from which the desirable fixed point cannot be reached and which are thus non-upstream. We consider the management option of lowering γ_2 to a value of, say, 1/2 by subsidising the clean technology to induce a technological change (Jaffe et al., 2002; Kalkuhl et al., 2012). This leads to the alternative dynamics depicted in Fig. 6, showing that for some initial states with $x_1 > x_2$ one can now get $x_2(t) \to \infty$ and $y_1(t) \to 0$. The goal of keeping the usage of the dirty technology below some limit, $y_1 < \ell < 1$, corresponds to a desirable region in terms of x_1, x_2, whose border can be computed as $x_2 = x_1(1/\ell - 1 - 1/\ell^{4/5}\sqrt{x_1})^{2/5}$ (see Fig. 6). That goal is automatically fulfilled in the top-left shelter region, can also be sustained by management (subsidies) in the glade region below it, and can at least be reached eventually from the remaining sunny upstream $U^{(+)}$ below the glade and from the dark upstream U^-, which is delimited by the management trajectory that meets the upper right corner.

But from below the latter trajectory, the shelter cannot be reached. In other words, when in U^-, one has to act fast in order not to lose the option of reaching S. From the dark part denoted Θ, not even the sunny region is reached, and hence that region is a trench, while the sunny part to its left is the abyss leading to that trench. There are no intermediate regions (downstream or eddies) between upstream and abyss in this example.

3.4 Combined population and resource dynamics

Our fourth example models the coevolution (in the sense of joint time evolution) of a natural Earth system component

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

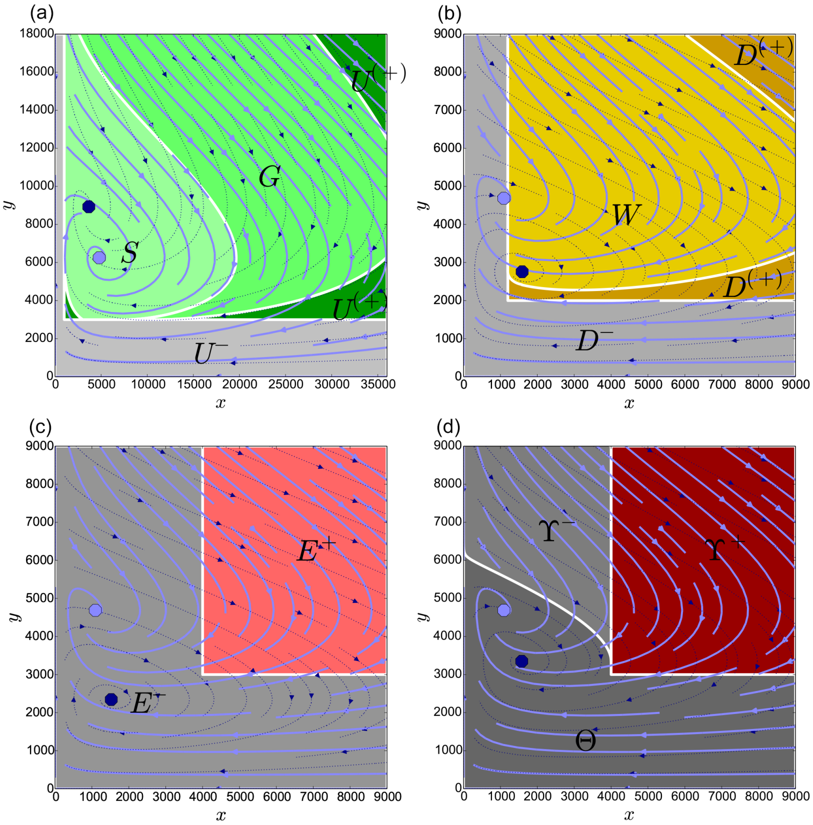


Figure 7. Combined population and resource dynamics. Coevolution of a population x and a resource stock y. In all cases, $\phi = 4$, $r = 0.04$. When the globally stable fixed point of the default dynamics (pale blue) falls into X^+, only upstream regions occur (top-left panel, $\gamma_0 = 4 \times 10^{-6} > \gamma_1 = 2.8 \times 10^{-6}$, $\delta = -0.1$, $\kappa = 12\,000$, $x_{\min} = 1000$, $y_{\min} = 3000$). When it falls into X^- instead, but the stable fixed point of the alternative management trajectory (dotted dark blue) is in X^+, then only downstream regions occur (top-right panel, $\gamma_0 = 8 \times 10^{-6} < \gamma_1 = 13.6 \times 10^{-6}$, $\delta = -0.15$, $\kappa = 6000$, $x_{\min} = 1200$, $y_{\min} = 2000$). Otherwise (bottom panels, $\gamma_0 = 8 \times 10^{-6} < \gamma_1$, $\delta = -0.15$, $\kappa = 6000$, $x_{\min} = 4000$, $y_{\min} = 3000$), the analysis depends on whether one can repeatedly reach X^+ by switching between default and alternative trajectories: for $\gamma_1 = 16 \times 10^{-6}$ (bottom-left panel), only eddies occur, while for $\gamma_1 = 11.2 \times 10^{-6}$ (bottom-right panel), only abysses and trenches occur.

coupled with a socio-economic Earth system component and shows how different parameters may qualitatively move the resulting state space topology through the whole main cascade, from an upstream-only situation via downstream-only and eddies-only to an abyss-and-trench situation.

The model was used in Brander and Taylor (1998) to explain the rise and fall of the native civilization on Rapa Nui (Easter Island) before western contact, but it may also be interpreted as a conceptual model of global population–vegetation interactions. It is derived from simple economic principles and leads to a modified Lotka–Volterra model with a finite resource. The human population x is preying on the island's forest stock y, which itself follows logistic growth dynamics:

$$\dot{x} = \delta x + \phi \gamma x y, \quad \dot{y} = r y (1 - y/\kappa) - \gamma x y$$

for some parameters γ, δ, κ, ϕ, and r representing growth and harvest rates and the stock's capacity.

We assume management will either reduce the default harvest rate γ_0 to some smaller value $\gamma_1 < \gamma_0$ to avoid overexploitation of the resource or increase it to a larger value $\gamma_1 > \gamma_0$ to avoid famine. Our choice of the sunny region relies on two principles. The absolute population should not drop below a threshold $x_{\min}$ and the relative decline in population under the default dynamics, $-\dot{x}/x$, should not exceed a value of ℓ. Hence $X^+ = \{x > x_{\min}$ and $y > y_{\min} = \max(0, -(\ell + \delta)/\phi \gamma_0)\}$.

The resulting state space partition is depicted in Fig. 7 for $\phi = 4$, $r = 0.04$ and different choices of γ_0, γ_1, δ, κ, $x_{\min}$, $y_{\min}$. One either gets an upstream-only situation, a downstream-only one, an eddy-only one, or an abyss-and-trench situation, depending on whether the unmanaged and

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

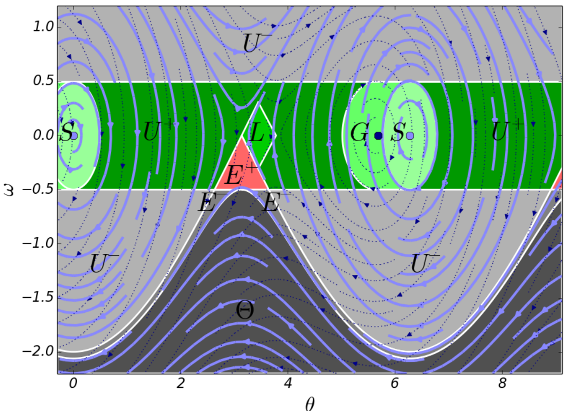


Figure 8. Gravity pendulum fun ride with management by one-sided acceleration and undesirable fast rotations. The 2π-periodic coordinate θ is the pendulum's inclination angle. If its angular velocity ω exceeds $\pm\ell$, people get sick (grey region). Since staying in L (balancing almost upright) or G (balancing somewhat inclined) is more exciting than in S (resting downward), we have both a glade and a lake dilemma.

managed fixed points belong to the desired or undesired region. In Appendix B2, these kinds of transitions are more formally interpreted as bifurcations.

An interesting case occurs when the whole state space is a single eddy as in Fig. 7 (bottom-left panel): one can then repeatedly visit the sunny region by suitably switching between a low default harvest rate and a managed higher harvest rate, but one cannot avoid getting back into the undesired region of a low or fast declining population. An "optimal" management strategy would then lead to slowly but strongly oscillating behaviour.

3.5 Gravity pendulum fun ride

While in the above examples typically only some of the possible regions were non-empty for each parameter combination, the following example from classical mechanics displays a rich diversity of state space regions that coexist at a single choice of parameter values. Despite extremely simple dynamics, it features both a glade and a lake dilemma, an eddy, and a trench at the same time.

In the model, people sit in a fun ride resembling a gravity pendulum with angle θ and angular velocity ω and default dynamics given by

$$\dot\theta = \omega, \quad \dot\omega = -\sin\theta.$$

An optional additional clockwise acceleration of the pendulum of magnitude $a > 0$ ("management") leads to alternative admissible trajectories on which for some time interval(s) one has $\dot\omega = -\sin\theta - a$. The sunny region is where $|\omega| < \ell$, for some $\ell > 0$ representing a safety speed limit above which people might get sick.

The unique shelter S is delimited by the default trajectory leading through the points $\theta = 2k\pi$, $\omega = \pm\ell$ that surrounds the stable resting state of $\theta = \omega = 0$ (see Fig. 8). If a state lies on a default trajectory that has $\omega > 0$ (anticlockwise pendulum motion) at least some of the time, then there is an admissible trajectory from it leading into the shelter, generated by the management strategy of "braking" whenever $\omega > 0$. Hence the upstream U equals the region strictly above the default trajectory with $\omega < 0$ that connects the unstable saddle point at $\theta = (2k+1)\pi$, $\omega = 0$ (pendulum balancing upright) with itself.

Just left of the shelter is the unique glade G. Depending on the parameter values, the stable fixed point of the managed dynamics (hanging pendulum inclined by constant acceleration) may either belong to the shelter or to the glade. In the latter case (Fig. 8), we have a glade dilemma since the inclined position is preferred to the resting position by the riders but is unsafe since if the engine breaks, people will get sick.

An even more exciting position is close to the upright balancing saddle point, at θ slightly larger than $(2k+1)\pi$ and $\omega \ll 1$, where there is an admissible trajectory that stays close to there (by braking repeatedly for short intervals while staying almost upright), so that this point is in the manageable region M. This is a typical example of how a region close to a saddle point of the default dynamics may become manageable due to an alternative feasible trajectory that has a slightly *shifted saddle point*, so that in the diamond-shaped region between the two saddle points, one can concatenate unmanaged and managed trajectories into periodic orbits.

However, for choices such as $a = 0.6$ and $\ell = 0.5$ (Fig. 8), there is no admissible trajectory leading from the exciting region with $\theta \approx (2k+1)\pi$, $\omega \approx 0$ into the shelter without entering the region with $|\omega| > \ell$. In that case the diamond-shaped region is a lake and we have a lake dilemma.

Finally, the region below and including the default trajectory that touches the line $\omega = -\ell$ from below is the trenches since one cannot brake in that direction, and the region between the trench and the upstream is the eddies. Downstream and abysses are empty in this example.

3.6 Bifurcations with manageable parameter

This final example system is designed to illustrate the relationship of reachability and bifurcations of a dynamical system that can be managed through a parameter and shows bifurcations of the type typically associated with tipping elements of the Earth system (Schellnhuber, 2009).

It has a two-dimensional state space $X = \{(r, y)\}$, where the "fast" variable $y \in \mathbb{R}$ has default dynamics

$$\dot y = h(y|r) = -\left(4+r^2\right)^3 y^3 + \left(2r^2-1\right)\left(4+r^2\right) y + e^r - 10,$$

which cannot be managed directly, and $r \in \mathbb{R}$ is a "slow" variable with (approximately) no default dynamics ($\dot r = 0$)

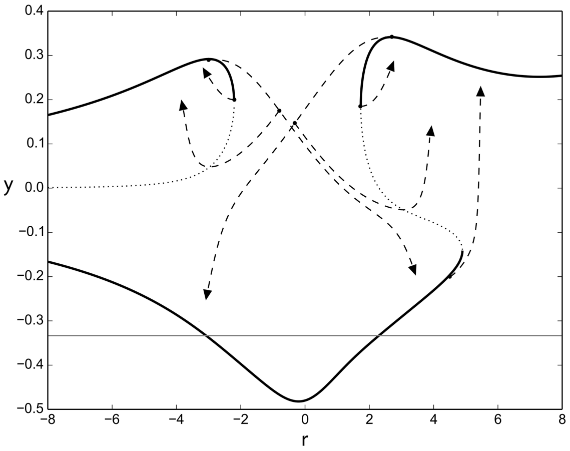


Figure 9. Bifurcations with manageable parameter. Loci of stable (solid black lines) and unstable (dotted lines) fixed points of $\dot{y} = -(4+r^2)^3 y^3 + (2r^2-1)(4+r^2)y + e^r - 10$. Leftmost and rightmost admissible management trajectories (dashed arrows) and their starting points (dots). Border (grey line) between sunny region $y > -1/3$ and the dark. See Fig. 10 for an analysis.

which, however, can be changed by management up to a velocity at most 100 and with arbitrarily large acceleration, leading to admissible trajectories with $\dot{r} \in [-100, 100]$ and $\dot{y} = h(y|r)$. We assume that values of $y \leqslant -1/3$ are undesirable.

If r is instead interpreted as a parameter of the one-dimensional system $\dot{y} = h(y|r)$, the set X can be interpreted as its bifurcation space in which one can plot a bifurcation diagram consisting of the loci of stable (solid lines) and unstable (dotted lines) fixed points, as shown in Fig. 9. As one can see, there are three saddle-node bifurcations at $r_1 \approx -2.2$, $r_2 \approx 1.735$, and $r_3 \approx 4.9$ with monostable parameter regimes $r_1 < r < r_2$ and $r > r_3$, and bistable parameter regimes $r < r_1$ and $r_2 < r < r_3$. Individual and paired saddle-node bifurcations (which often result from fold bifurcations) occur frequently in bistable Earth system components such as the hysteretic thermohaline circulation (Stommel, 1961; Rahmstorf et al., 2005), monsoonal soil–vegetation feedbacks (Janssen et al., 2008), or other tipping elements (Schellnhuber, 2009). Hysteresis also occurs on other spatial and temporal scales, e.g. in local hydrology (Beven, 2006) and in long-term glacial climate dynamics (Ganopolski and Rahmstorf, 2001).

The main part of the resulting network of ports and rapids of our example system is depicted in Fig. 10. On its coarsest level, there are two ports, each containing one of the two connected loci of stable/unstable fixed points, and a rapid in between through which one can pass from the left to the right port but not back. If the right port seems more attractive, e.g. because it allows a higher value of y, we have a port dilemma since by leaving the left port for the right one, we lose flexibility in terms of reachable regions.

The right port contains two harbours, similarly connected by a narrow "internal" channel, as well as another "exit" channel leading from the right harbour to the dark region. Note that on the leftward-pointing dashed management trajectory in the middle of the bifurcation diagram, there is a leftmost point from where one can still "turn around" and reach (if only unstably) the right part without entering the dark region; this point is a corner of the right harbour (but not belonging to it, for stability reasons), and below it is a channel leading to another harbour in the bottom left. Again, if the right harbour seems more attractive, we have a dilemma, this time a harbour dilemma, since in order to reach the right harbour from the left one, we have to pass through the dark.

Finally, the right harbour contains two docks again connected by a fairway, plus some more fairways. Again, we get a dilemma if the top-right dock is more attractive than the top-left one: the dock dilemma is that, in order to reach the top-right dock from the top-left one, one has to pass through the unsafe middle region and risk ending up in the dark if management breaks down.

4 Discussion and conclusions

We have presented a formal classification of the possible states of a dynamical system such as the Earth system into regions of state space which differ qualitatively in their safety, the possibilities of reaching a safe state, the possibilities of avoiding undesired states, and in the amount of flexibility for future management.

Based on an assumed main division of the system's states into only two classes, desirable ("sunny") and undesirable ("dark"), we have constructed a hierarchy of partitions of a system's state space, whose member regions we suggested to name by metaphorical names either corresponding to the general image of a boat floating or rowing on a complex water system, such as "upstream", "downstream", "eddy", "abyss", "trench", "lake", and "backwater", or corresponding to the image of a "shelter" surrounded by a "glade". To capture the nature of and relationships between the different regions, we have introduced the notion of stable reachability and the corresponding three-level reachability network of "ports", "harbours", "docks", "rapids", "channels", and "fairways", and illustrated our concepts with conceptual example models from climate science, ecology, coevolutionary Earth system modelling, economics, and classical mechanics. Most of the different regions can readily be found in most models for either most or at least selected parameter settings. A notable exception is the "eddies", which, due to their circular nature, can be expected to occur much more rarely in real-world, non-conservative systems, especially when thermodynamic or otherwise irreversible processes are involved, such as soil degradation. Section 3.4, however, illustrates how eddies may occur in coevolutionary systems and might incentivize management cycles that lead to undampened periodic

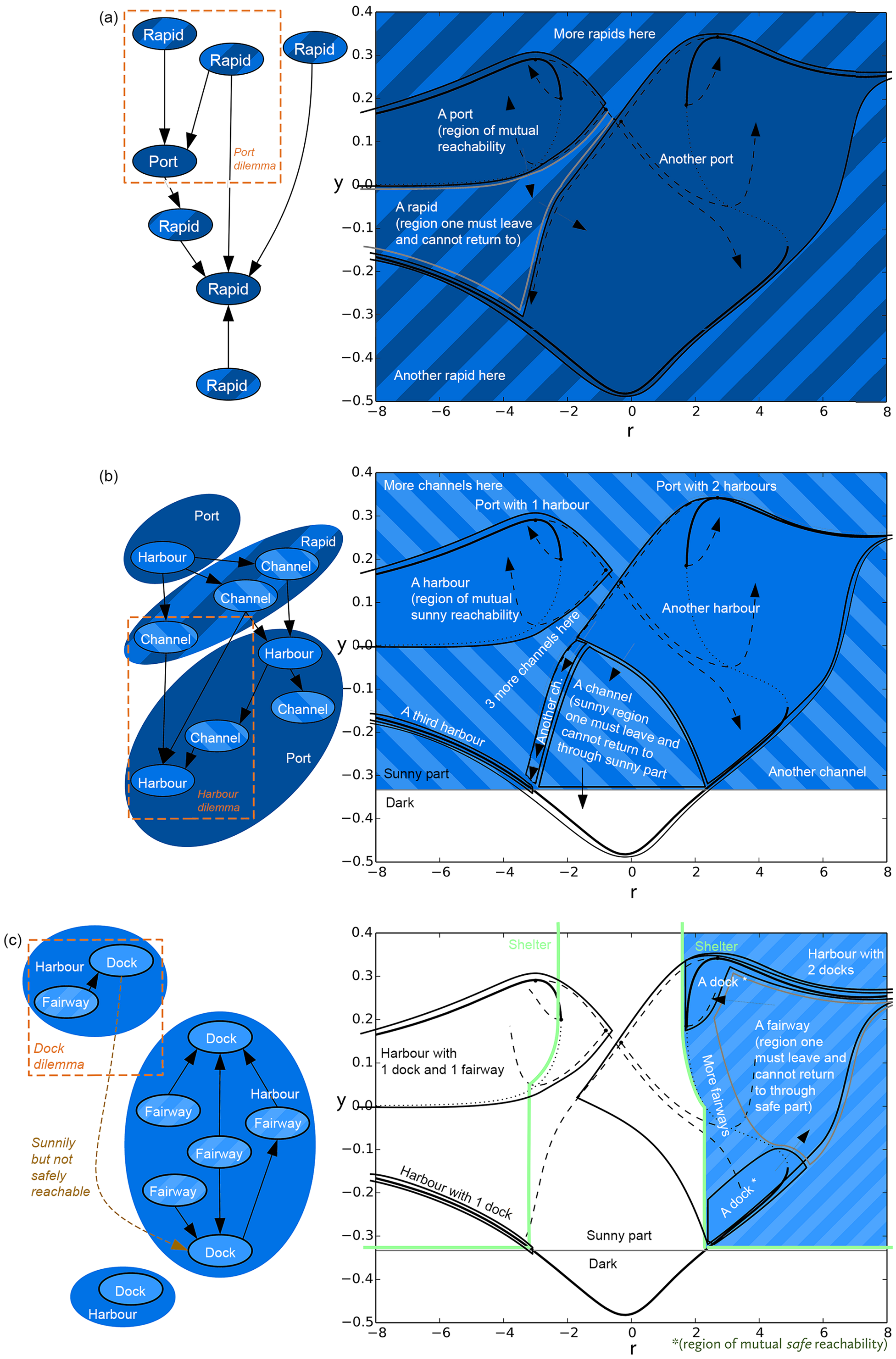


Figure 10. Main part of the three-level reachability network of ports and rapids (top panel), harbours and channels (middle panel), and docks and fairways (bottom panel, and related dilemmas in the bifurcation example. Arrows indicate stable reachability (top panel), stable reachability through the sun (middle panel), and stable reachability through the shelters (bottom panel). Some further arrows between rapids, channels and fairways have been omitted here.

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

ups and downs. It must remain an open question here whether this effect might be an additional explanation for empirically observable cycles such as business or resource cycles when management is involved.

The introduced concepts have then been used to point out a number of qualitatively different decision problems: the glade, lake, port, harbour, and dock dilemmas. In our opinion, one particularly nasty form of decision problem is the lake dilemma, where one has to choose between uninterrupted desirability and eventual safety, and Sect. 3.2 indicates that this dilemma may easily occur at least in ecological systems or other multistable systems with a sunny attractor and another one slightly in the dark. Since the transformation of socio-metabolic processes or complex industrial production systems may resemble the soil transformation of Sect. 3.2, one may also expect the lake dilemma to occur in the socio-metabolic and economic subsystems of the Earth, e.g. in the context of a great transformation leading to decarbonisation of the world's energy system. The form of lake seen near the saddle point in the pendulum (Sect. 3.5) can also occur in other nonlinear oscillators, e.g. the Duffing oscillator or models of glacial cycles that resemble it such as Saltzman et al. (1982) and Nicolis (1987), when a management option exists that has a slightly shifted saddle point. This indicates that the lake dilemma may also occur in purely physical subsystems of the Earth system.

We argue that our concepts may be especially useful in the context of the current debate about planetary boundaries (PBs), a possible safe and just operating space (SAJOS) for humanity, and the necessary socio-economic transitions to reach it or stay in it. We suggest that the region delimited by some identified set of PBs in the sense of Rockström et al. (2009a) and Steffen et al. (2015) and some similar socio-economic limits, e.g. those relating to the United Nations sustainable development goals (Raworth, 2012), should be interpreted in our framework as a natural choice for the desirable region X^+, although their definitions already contain some reasoning about the consequences for the respective subsystems when the boundaries are violated. Such boundaries might be called the ultimate planetary boundaries (UPBs), and they are typically defined by some simple thresholds for relevant indicators as in Rockström et al. (2009a) and Steffen et al. (2015), not taking into account the *overall* system's inherent dynamics much. In this sense, UPBs are typically "non-interacting". Based on the UPBs, one may then try to identify one or more smaller shelter regions S that can be considered a SAJOS in the sense that, once there, no further large-scale management in the form of global policies is necessary to stay within the limits for all times (or at least for a sufficiently long planning horizon). The borders of these shelters are also a form of PBs but are much more restrictive than the UPBs we started with, and we suggest to call them safe planetary boundaries (SPBs).

If it turns out that the current state of the Earth is outside the shelters, one should then aim next at trying to decide whether it is in the upstream. If so, knowledge about whether it is in a glade or lake or not, and which safe docks can be stably reached, will be necessary in order to choose a management path. In the glade case, one can still reach the shelter without ever violating the UPBs by appropriate management; hence we suggest to refer to the border of shelters and glades together as the provident planetary boundaries (PPBs).

In the lake case, one has to decide instead whether a temporary violation of the UPBs can be justified by the eventual safety of the shelters. In addition, a port dilemma may necessitate a decision between higher desirability and higher flexibility at this point. Only after these qualitative decisions have been made does it seem advisable to optimize the chosen type of management pathway by means of more traditional control and optimization theory, hopefully using accurate enough quantitative estimates of the involved options, costs, and benefits. Once in the shelters, one may start caring about improving the state further by moving between docks to either improve desirability or flexibility, but this may require a risky temporary passage through a sunny but unsafe region (which poses a dock dilemma) or even a passage trough the dark (which poses a harbour dilemma). Of course, many combinations of these qualitative and quantitative criteria may appear in the actual global decision process, e.g. in the form of lexicographic preferences, decision trees, or more sophisticated welfare measures or other quantitative objective functions that take the topology suitably into account and that may relate to some form of market (or other game-theoretic) equilibrium or else be governed by some suitable policy instruments, as kindly suggested by an anonymous referee.

If we are not in the "upstream" of the Earth system, prospects are worse. Violating the limits can then only be avoided by management, either eventually forever (if in the downstream), or only repeatedly but with repeated violations occurring (if in the eddies), or even only for a limited time with an ultimate descent into the undesired region (if in the abysses or already in the trench). We suggest to call the upstream borders the no-regrets planetary boundaries (NRPBs).

If the diagnosis reads "eddy", "abyss", or "trench", one may repeat the analysis with a less ambitious, "second best" definition of the desirable region by choosing less restrictive UPBs, or revert to quantitative optimization, e.g. to minimize some damage function along the system's trajectory. On the other hand, as long as one is in the "manageable region" M (shelters, glades, lakes, and backwaters), the UPBs need never be transgressed if managed wisely; hence we propose to call the borders of M the foresighted planetary boundaries (FPBs).

This completes our suggested hierarchy of PBs from the relatively looser UPBs via the successively narrower FPBs and NRPBs, then the PPBs, to the narrowest SPBs that define the SAJOS. While UPBs are "non-interacting", FPBs, PPBs, NRPBs, and SPBs will typically have a more complex geometry in the system's state space and are thus "interacting boundaries". This means that they cannot be expressed as a simple "threshold" for individual indicators but as conditional thresholds for several indicators that depend on each other as shown by the curved region boundaries in the examples, e.g. in the carbon cycle model of Anderies et al. (2013) in Sect. 3.1. Obviously, the real world is less black and white than suggested by the idealized division into "desirable" and "undesirable", so the actual location of these bounds will in reality be somewhat vague, but this does not change the fact that the different bounds and regions represent qualitatively different states of the system, not just quantitative shades of grey.

It should be noted that one strategy to decide the dilemmas described throughout this work is to follow certain "sustainability paradigms" such as those suggested by Schellnhuber (1998). For example, the "pessimization paradigm" is based on the basic precautionary principle of "avoiding the worst" and, hence, can be interpreted as suggesting to stay in or aim for the shelter. In this way, the "pessimization paradigm" decides the glade and lake dilemmas in favour of safety. In turn, the "optimization paradigm" could be interpreted to decide all but the harbour dilemma in favour of uninterrupted or (eventually) higher desirability. The "stabilization paradigm", which seems to fit best the popular notions of "sustainable development", reflecting a "longing for stable equilibria" in the coevolutionary dynamics of human societies and the biophysical Earth system (Schellnhuber, 1998), might imply staying in a lake favouring uninterrupted desirability over eventual safety in the sense of this work. Finally, the "equitization paradigm" might imply choosing higher flexibility, e.g. in terms of a larger set of remaining options for future generations in the sense of intergenerational justice, in all dilemmas but the lake dilemma. As also argued by Schellnhuber (1998), the remaining "standardization paradigm" is entirely based on static choices of norms or development corridors instead of dynamical systems or "geocybernetic" principles and, hence, cannot directly decide any of the dilemmas. However, this paradigm can be viewed as a way for identifying desirable domains in the Earth system's state space in the first place and, thereby, facilitate a subsequent topological classification of state space structure.

Contemplating sustainability paradigms gives rise to other relevant qualitative decision problems. For what might be called an "optimization/pessimization dilemma", consider the debate on geoengineering by solar radiation management (Lenton and Vaughan, 2009; Vaughan and Lenton, 2011) as a strategy for averting some of the consequences of global climate change that are induced by anthropogenic emissions of greenhouse gases (Stocker et al., 2013). According to the recent update of the planetary boundary framework by Steffen et al. (2015) and the corresponding definition of desirability (see Sect. 1.1, "Metaphorical framework"), the Earth system is currently in the dark region of its state space, because core planetary boundaries such as those related to climate change and biosphere integrity have likely already been transgressed. Following current assumptions on the feasibility of management options (IPCC, 2014), assume further that the Earth system is currently in the dark upstream. In this situation, efforts for mitigation of greenhouse gas emissions, e.g. by means of global energy market regulations, as well as conservation and restoration of biosphere integrity, would correspond to navigating the Earth system from the dark upstream towards the shelters following the "pessimization paradigm". In turn, massive investments in solar radiation management as an alternative to mitigation could be seen as manoeuvring the Earth system into the glades or lakes going along with a severe loss of resilience, since interruption of these efforts due to global crisis or technological failure would lead to very rapid and catastrophic climate change (Barrett et al., 2014). In short, starting in the dark upstream, does one choose to navigate to a glade or lake because this appears economically cheaper on the shorter term or politically more feasible ("optimization paradigm") or does one aim for the shelters right away, even if this is more expensive on the shorter term ("pessimization paradigm")? Note, however, that geoengineered Earth system states within the glades or lakes would be expected to have a considerably reduced desirably in the long-term compared to the shelters, since current proposals for solar radiation management can only control a very small set of Earth system properties such as global mean temperature, while regional temperature patterns and the hydrological cycle would change strongly (Kleidon and Renner, 2013; Kleidon et al., 2015), going along with corresponding climate impacts.

We hope that the theoretical considerations outlined here may be of some help to sharpen the important debate of how a transition to a safe desirable state of the Earth system can be managed. To this end, future studies should apply the proposed framework for comparing different Earth system governance strategies in the form of various management options (e.g. mitigation of greenhouse gas emissions vs. geoengineering) and different notions of desirability (e.g. resemblance of a Holocene-like state or satisfaction of a certain standard of human well-being) in terms of their feasibility and resilience. Furthermore, the structural stability of future development pathways generated by integrated assessment models through optimizing utility functions based on certain notions of human well-being could be evaluated. For achieving these aims, performant computer algorithms need to be developed for automatically generating the proposed topological charts also for higher-dimensional Earth system models given a set of management options and desirability criteria, e.g. building on algorithms from viability theory (Frankowska and Quincampoix, 1990), the graph-theoretical analysis of phase space transition networks (Padberg et al., 2009), and flow networks from fluid dynamics (Ser-Giacomi et al., 2015; Froyland and Padberg-Gehle, 2015). While the examples discussed in this work have been limited to two dynamical variables for facilitating the visualization of the corresponding topological charts, investigation of more detailed

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

models of Earth system dynamics calls for advanced visualization techniques (Nocke et al., 2015) as well as the application and further development of quantitative measures of the size (Menck et al., 2013; Hellmann et al., 2015; van Kan et al., 2015) and shape (Mitra et al., 2015) of the phase space regions of interest. The fact that the introduced state space partitions depend on qualitative rather than quantitative properties of states may also make them a natural tool for the analysis of complex but qualitative or "generalized" models in the spirit of Kuipers (1994) and Petschel-Held et al. (1999) or Lade et al. (2013, 2015b, a).

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

Appendix A: Formal derivation of partitions and properties

We use sloppy set theoretic notation when no confusion arises: union $A+B=A\cup B$, difference $A-B=A\setminus B$, power set $2^A=\{B\subseteq A\}$. Proofs only require an understanding of general topological spaces, in particular of openness and continuity, but not of any higher-level concepts from differential topology or the like.

A1 Assumptions and notation

For a more formal treatment than in the main text, we assume a *manageable dynamical system with desirable states*, made of the following ingredients.

A *state space* $X\neq 0$ with some Hausdorff topology $\mathcal{T}\subseteq 2^X$ (i.e. a system of open sets that separate each two points) on it whose elements we call *states* or *points* (e.g. $X\subseteq \mathbb{R}^n$ with Euclidean topology). X may be compact or unbounded, finite- or infinite-dimensional, etc.

A flow (i.e. deterministic continuous-time autonomous dynamical system) on X (e.g. a model of human-nature coevolution or any other Earth system model) given by a family of continuous ("business-as-usual" or) *default trajectories* $\tau_x:[0,\infty)\to X$ with $\tau_x(0)=x$ and $\tau_{\tau_x(t)}(t')=\tau_x(t+t')$ for all initial conditions $x\in X$ and all relative time points t, $t'\geqslant 0$. We do not require further smoothness properties of the flow, like differentiability, to avoid having to assume a richer topological structure for X than just a general topological space, and to avoid unnecessarily complicated notions and familiarity with, for example, differential geometry. Although flows are often represented by ordinary differential equations, their solutions are sometimes not unique, and hence our notion of flow is in terms of trajectories instead so as to allow us to distinguish, for example, a 1-D flow with $\dot{x}=\sqrt{x}$ and $\tau_0(t)\equiv 0$ from the flow that also has $\dot{x}=\sqrt{x}$ but $\tau_0(t)=t^2/4$.

An open nonempty set $X^+\in\mathcal{T}$ of desirable states, called the *sunny region*, e.g. defined by means of some notion of "tolerable E & D window" (Schellnhuber, 1998). We call the complement $X^-=X-X^+\neq 0$ the *dark (region)*. We require openness for convenience so that infinitesimal perturbations cannot lead from the sunny to dark part, and trajectories cannot touch the sunny region without entering it for a strictly positive amount of time. Although in most of our examples, X^+ is a simply shaped, connected, convex, and often bounded set, none of these properties is required for the theory presented here except topological openness.

To represent "management options", a family of nonempty sets $\mathcal{M}_x$ of *admissible trajectories* from each $x\in X$ that includes τ_x and is closed under switching between trajectories at any time, i.e. if $\mu\in\mathcal{M}_x$, $t>0$, $x'=\mu(t)$, and $\mu'\in\mathcal{M}_{x'}$, then the trajectory defined by $\mu''(t'')=\mu(t)$ for $t''\leqslant t$ and $\mu''(t'')=\mu'(t''-t)$ for $t''>t$ is also in $\mathcal{M}_x$. This requirement corresponds to the so-called semigroup axiom of mathematical control theory (Sontag, 1998). Note that we do not allow any explicit time dependency of flow or management, but such dependencies can as usual be encoded by including time as a state variable. Also, if management can change a parameter of the model, that parameter has to be transformed to a (slow) state variable with zero default dynamics of its own to meet our framework.

A2 Open invariance, sustainability, and stable reachability

The *invariant open kernel* of a set $A\subseteq X$, denoted $A^{\circlearrowleft}$, is the largest open subset of A that contains the default trajectories of all its own points. Its existence and uniqueness is nontrivial and will be proved below. Note that $A^{\circlearrowleft}$ may be empty. Each (topologically) connected component of $S=(X^+)^{\circlearrowleft}$ is called an individual *shelter*.

We call an open set $A\in\mathcal{T}$ *sustainable* iff, for all $x\in A$, there is $\mu\in\mathcal{M}_x$ with $\mu(t)\in A$ for all $t\geqslant 0$. Again, the openness requirement ensures a minimal form of stability against small perturbations. The *sustainable kernel* of a set $A\subseteq X$, denoted $A^{\mathcal{S}}$, is the largest sustainable open subset of A. Again, existence and uniqueness will be proved below. In viability theory (Aubin, 2001), $A^{\mathcal{S}}$ roughly corresponds to the "viability kernel" of A (see the discussion in Supplement 3). Also, $A^{\mathcal{S}}$ may be empty.

Lemma 1 (Existence and uniqueness) *For all* $A\subseteq X$:

1. *There is a unique largest (default-trajectory-) invariant and open subset* $A^{\circlearrowleft}\subseteq A$, *containing all other such sets.*
2. *Every invariant and open set is sustainable. In particular,* S *is.*
3. *There is a unique largest sustainable subset* $A^{\mathcal{S}}\subseteq A$ with $A^{\mathcal{S}}\supseteq A^{\circlearrowleft}$, *containing all other such sets.*

Proof.

1. Let $\mathcal{I}(A)$ be the system of all open subsets $B\subseteq A$ for which $\tau_x(t)\in B$ for all $x\in B$, $t>0$. The proposition is proved by showing that $\mathcal{I}(A)$ is a *kernel system*, i.e. contains the empty set (which is trivial) and contains the union $\bigcup\mathcal{B}$ of any of its subsets $\mathcal{B}\subseteq\mathcal{I}(A)$. The latter follows from the fact that the system of all open sets, $\mathcal{T}$, is a kernel system by definition, and if $x\in\bigcup\mathcal{B}$, then $x\in B\in\mathcal{B}$, and hence $\tau_x(t)\in B\subseteq\bigcup\mathcal{B}$ for all $t>0$. Now $A^{\circlearrowleft}=\bigcup\mathcal{I}(A)\in\mathcal{I}(A)$.
2. This follows because $\tau_x\in\mathcal{M}_x$.
3. Similarly, the system $\mathcal{S}(A)$ of all sustainable subsets $B\subseteq A$ is a kernel system: if $x\in\bigcup\mathcal{B}$, then $x\in B\in\mathcal{B}$, and hence there is $\mu\in\mathcal{M}_x$ with $\mu(t)\in B\subseteq\bigcup\mathcal{B}$ for all $t>0$. Now $A^{\mathcal{S}}=\bigcup\mathcal{S}(A)\in\mathcal{S}(A)$. Point 2 implies $A^{\mathcal{S}}\supseteq A^{\circlearrowleft}$.

Q. E. D.

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

Next, we introduce a suitable notion of stable reachability to overcome two problems with the classical notion of (plain) reachability known from control theory, where a state y is reachable from another state x iff it lies on some admissible trajectory starting at x (Sontag, 1998).

First, we want a stable fixed point y of the default dynamics to be counted as stably reachable from a (sufficiently small) neighbourhood of itself, although one might only get arbitrarily close to y instead of getting to y in finite time. Second, we want stable reachability to imply that small perturbations along the way cannot render the target unreachable. To solve this conceptual task in a mathematically convenient way, we define stable reachability here via the following binary relation between sets. We call an open set $C \in \mathcal{T}$ a *forecourt* for some set $Y \subseteq X$, denoted $C \rightsquigarrow Y$, iff one can approach Y arbitrarily closely from everywhere in C without leaving C, or, more precisely, iff for all $x \in C$, there is $\mu \in \mathcal{M}_x$ so that, for all open sets $Z \in \mathcal{T}$ with $Z \supseteq Y$, there is $t > 0$ with $\mu(t) \in Z$ and $\mu(t') \in C$ for all $t' \in [0, t]$. Now, for a state $x \in X$ and some set $A \subseteq X$, we say that another state $y \in X$ or another set $Y \subseteq X$ is *stably reachable from x through A*, denoted $x \rightsquigarrow_A y$ or $x \rightsquigarrow_A Y$, iff x is in some subset of A that is a forecourt for $\{y\}$ or Y, respectively. The set of states from where Y can be stably reached through A is denoted $(\rightsquigarrow_A Y)$. (This is a stable version of what Aubin, 2001, would call a "capture basin" of Y.) Note that in these definitions, the order in which the logical quantifiers "for all" and "there exists" appear is critical for some of the resulting properties. If Y is open, the definitions can be somewhat simplified:

Proposition 1 (Stable reachability)
For all $A, A', C, Y, Z \subseteq X$ and $x, y, z \in X$:

1. *If Y is open, then* (i) *$C \rightsquigarrow Y$ iff, for all $x \in C$, there is $\mu \in \mathcal{M}_x$ so that there is $t > 0$ with $\mu(t) \in Y$ and $\mu(t') \in C$ for all $t' \in [0, t]$, and (ii) $x \rightsquigarrow_A Y$ iff there is and open $C \subseteq A$ with $x \in C$ and for all $x' \in C$, there is $\mu \in \mathcal{M}_{x'}$ so that there is $t > 0$ with $\mu(t) \in Y$* and *$\mu(t') \in C$ for all $t' \in [0, t]$.*

2. *If $x \rightsquigarrow_A Y$, then x is in the interior* (i.e. *largest open subset*) *of A, A°, and there is an open set $B \ni x$* with *$x' \rightsquigarrow_A Y$ for all $x' \in B$. Hence, each set of the form $(\rightsquigarrow_A Y)$ is open.*

3. *Transitivity*:

   $$x \rightsquigarrow_A y \rightsquigarrow_{A'} Z \Longrightarrow x \rightsquigarrow_{A+A'} Z,$$
   $$x \rightsquigarrow_A y \rightsquigarrow_{A'} z \Longrightarrow x \rightsquigarrow_{A+A'} z.$$

   In particular, $\rightsquigarrow_A$ is a transitive (but not necessarily reflexive) relation.

4. *If A is open, it is stably reachable from each of its elements. In particular, since $S = (X^+)^{\wp} \subseteq (X^+)^{\mathcal{S}} = M$ is open, S is also included in $U = (\rightsquigarrow_X S)$.*

Proof.

1. (i) Assume $C \rightsquigarrow Y \in \mathcal{T}$ and let $x \in C$. Then, by definition of forecourts, there is $\mu \in \mathcal{M}_x$ so that, for all open sets $Z \in \mathcal{T}$ with $Z \supseteq Y$, there is $t > 0$ with $\mu(t) \in Z$ and $\mu(t') \in C$ for all $t' \in [0, t]$. Since Y is open, it is such a Z, proving the first direction.

   For the other direction, assume that for all $x \in C$, there is $\mu \in \mathcal{M}_x$ so that there is $t > 0$ with $\mu(t) \in Y$ and $\mu(t') \in C$ for all $t' \in [0, t]$. Let $x \in C$, choose such a $\mu \in \mathcal{M}_x$ and $t > 0$, and let $Z \in \mathcal{T}$ with $Z \supseteq Y$ be an open set. Then $\mu(t) \in Y \subseteq Z$ as required.

   (ii) By definition of stable reachability, $x \rightsquigarrow_A Y$ iff there is an open $B \subseteq A$ with $x \in B \rightsquigarrow Y$. By (i), $B \rightsquigarrow Y$ iff for all $x' \in B$, there is $\mu \in \mathcal{M}_{x'}$ so that there is $t > 0$ with $\mu(t) \in Y$ and $\mu(t') \in B$ for all $t' \in [0, t]$.

2. Assume $x \rightsquigarrow_A Y$. Then $x \in X$ for some open $B \subseteq A$, and hence $x \in B \subseteq A^\circ$. Also, $B \rightsquigarrow Y$ and hence $x' \rightsquigarrow_A Y$ for all $x' \in B$. Hence $(\rightsquigarrow_A Y)$ contains an open neighbourhood of each of its points and is thus open itself.

3. We show this by concatenating suitably chosen admissible trajectories between points close to x, y, Z. Let $x \rightsquigarrow_A y \rightsquigarrow_{A'} Z$, choose open sets $B \subseteq A$, $B' \subseteq A'$ with $x \in B \rightsquigarrow \{y\}$ and $y \in B' \rightsquigarrow Z$, and put $B'' = B + B' \subseteq A + A'$, then $x \in B''$ and B'' is open. To show that $B'' \rightsquigarrow Z$, we let $x'' \in B''$ and show that there is $\mu \in \mathcal{M}_{x''}$ so that, for all open $W'' \supseteq Z$, there is $t > 0$ with $\mu(t) \in W''$ and $\mu(t') \in B''$ for all $t' \in [0, t]$.

   If $x'' \in B'$, there is such a μ with $\mu(t') \in B' \subseteq B''$ for all $t' \in [0, t]$ since $B' \rightsquigarrow Z$.

   If $x'' \notin B'$ instead, $x'' \in B \rightsquigarrow \{y\}$, and hence we find $\nu \in \mathcal{M}_{x''}$ so that, for all open $W \supseteq \{y\}$, there is $t > 0$ with $\nu(t) \in W$ and $\nu(t') \in B$ for all $t' \in [0, t]$. Since B' is such a W, we find $t'' > 0$ with $\nu(t'') \in B'$ and $\nu(t') \in B$ for all $t' \in [0, t'']$. For $y' = \nu(t'') \in B' \rightsquigarrow Z$, we then find $\nu' \in \mathcal{M}_{x''}$ so that, for all open $W'' \supseteq Z$, there is $t > 0$ with $\nu'(t) \in W''$ and $\nu'(t') \in B'$ for all $t' \in [0, t]$. Now define μ by putting $\mu(t') = \nu(t')$ for $t' \in [0, t'']$ and $\mu(t') = \nu'(t' - t'')$ for $t' \geqslant t''$. Then $\mu \in \mathcal{M}_{x''}$ because of our assumptions on $\mathcal{M}$, and for all open $W'' \supseteq Z$, there is $t > 0$ with $\nu'(t) \in W''$ and $\nu'(t') \in B + B' = B''$ for all $t' \in [0, t]$, as required.

   The z case follows from putting $Z = \{z\}$. Transitivity is the special case of $A' = A$.

4. For $x \in A \in \mathcal{T}$, we show $x \rightsquigarrow_A A$ by showing $A \rightsquigarrow A$. Let $x' \in A$. By (1), we have to find $\mu \in \mathcal{M}_{x'}$ and $t > 0$ with $\mu(t') \in A$ for all $t' \in [0, t]$. Since A is open and $\tau_{x'}$ is continuous, $\tau_{x'}$ is such a μ.

Q. E. D.

A3 Partitions

A topologically connected component of

$$\Theta = X - \left(\rightsquigarrow_X X^+\right),$$
$$\Upsilon = \overline{\{x \in X | \forall \mu \in \mathcal{M}_x \exists t \geqslant 0 : \mu(t) \in \Theta\}} - \Theta,$$

or

$$E = X - U - D - \Upsilon - \Theta$$

will be called an individual *trench*, *abyss*, or *eddy*, and the latter two typically have sunny and dark parts. Some further properties of these introduced partition sets are as follows.

Proposition 2 (Main cascade).

1. *$U = (\rightsquigarrow_X S)$ and the union $D + U = (\rightsquigarrow_X M)$ are open, $\Theta = X - (\rightsquigarrow_X X^+)$ and $\Upsilon + \Theta$ are closed, the union $E + D + U = X - \Upsilon - \Theta$ is open, and the system $\{U, D, E, \Upsilon, \Theta\}$ forms a partition of X.*
2. *For all $u \in U$, $d \in D$, $e \in E$, $y \in \Upsilon$, $\theta \in \Theta$, we have $\neg(\theta \rightsquigarrow_X y)$, $\neg(y \rightsquigarrow_X e)$, $\neg(e \rightsquigarrow_X d)$, $\neg(d \rightsquigarrow_X u)$.*
3. *If $W = \varnothing$, also $D = \varnothing$.*

Proof.

1. Openness follows from Proposition 1, the partition covers X by definition of E, and the only nontrivial disjointness is that between the open set $D + U = (\rightsquigarrow_X M)$ and the closed set $\Upsilon + \Theta = \overline{\{x \in X | \forall \mu \in \mathcal{M}_x \exists t \geqslant 0 : \mu(t) \in \Theta\}}$. If x is in both sets, there is also $x' \in (\rightsquigarrow_X M) \cap \{x \in X | \forall \mu \in \mathcal{M}_x \exists t \geqslant 0 : \mu(t) \in \Theta\}$, but then there is $\mu'_x \in \mathcal{M}_x$, $t' > 0$ with $\mu'_x(t') \in M$, and by definition of M there is then also some $\mu \in \mathcal{M}_x$ with $\mu(t) \in X^+$ for all $t \geqslant t'$. But, by assumption, there is $t \geqslant 0$ with $\mu(t) \in \Theta$. Since $\Theta \cap X^+ = 0$, we have $t < t'$, but by definition of Θ, this contradicts $\mu(t') \in X^+$. Hence such an x cannot exist.
2. Because of transitivity and (1), $d \rightsquigarrow_X u \in U = (\rightsquigarrow_X S)$ would imply $d \rightsquigarrow_X S$ and thus $d \in U \cap D = \varnothing$; $e \rightsquigarrow_X d \in D = (\rightsquigarrow_X M) - U$ would imply $e \rightsquigarrow_X M$ and thus $e \in (U + D) \cap E = \varnothing$. If one could reach the eddies from the abysses, one could avoid the trenches: assume $y \rightsquigarrow_X e \notin \Upsilon + \Theta = \overline{\{x \in X | \forall \mu \in \mathcal{M}_x \exists t \geqslant 0 : \mu(t) \in \Theta\}}$. Since the latter is closed, its complement is open, so there is $\mu \in \mathcal{M}_y$ and $t > 0$ with $\mu(t) \notin \Upsilon + \Theta$. For $x = \mu(t)$, we find $\mu' \in \mathcal{M}_x$ and $t'' > 0$ with $\mu'(t') \notin \Theta$ for all $t' > t''$. Concatenating μ with μ' gives a similar member of $\mathcal{M}_y$, in contradiction to $y \in \Theta$. Finally, if $\theta \rightsquigarrow_X y$ and $\theta \in \Theta$, then $y \in \Theta$ by definition of Θ, and hence $y \notin \Upsilon$.
3. This follows from $(\rightsquigarrow_X M) - U = D = (\rightsquigarrow_X W)$.

Q. E. D.

Note that in the (pathological) *no-management case* in which $\mathcal{M}_x = \{\tau_x\}$, the upstream $U = (\rightsquigarrow_X S)$ is basically (i.e. up to boundary effects due to our openness requirement) the basin of attraction of S, the downstream $D = (\rightsquigarrow_X M) - (\rightsquigarrow_X S)$ is then empty, the trenches basically equal the invariant kernel of X^-, the abysses basically equal the rest of the basin of attraction of the trenches, and the eddies are basically the union of those trajectories that will forever alternate between X^+ and X^-. In that case, some of the finer regions may coincide or be empty as well, and one can also represent their relationship by means of *symbolic dynamics* (beim Graben and Kurths, 2003): assign each state x a symbolic sequence representing the sequence of its trajectory's transitions between the sunny $(+)$ and dark $(-)$ regions, and use the wildcard $*$ to denote repetitions of zero or more symbols. Then (up to peculiarities that may occur for boundary states)

$$S = M = (+),$$
$$U^- = (-)(+-)^*(+),$$
$$U^{(+)} = (+-)(+-)^*(+),$$
$$G = L = D = \varnothing,$$
$$E^+ = (+-)^\infty,$$
$$E^- = (-+)^\infty.$$
$$\Upsilon^+ = (+)(-+)^*(-),$$
$$\Upsilon^- = (-+)(-+)^*(-),$$

and

$$\Theta = (-).$$

To formally define the ports-and-rapids partition, we say that a set $P \subseteq X$ is *portish* iff it has $x \rightsquigarrow_X y$ for all $x, y \in P$; is topologically connected; and does not intersect two different eddies, abysses, or trenches. A maximal portish set is called a *port*.

We show below that all ports are disjoint; each port is completely contained in one of the sets U, D, E, Υ^-, Θ; none can intersect Υ^+; and each *returnable* state (i.e. an x with $x \rightsquigarrow_X x$) is in a port, but no *transitional* state (x with $\neg(x \rightsquigarrow_X x)$) is.

In the pendulum example of Fig. 8, the returnable points are those in $U + D$ because of the periodic frictionless default flow and the possibility of counteracting small perturbations by braking or acceleration at some later point of the perturbed trajectory. In the eddies and below, this is not possible after an accelerating perturbation; hence those regions are transitional. In the plant types example of Fig. 5, there are also transitional regions, e.g. to the top and right, where all admissible trajectories lead down and left, and in the technological change example of Fig. 6, all points are transitional because of the positive growth of the knowledge stocks.

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

To extend the system $\mathcal{P}$ of all ports into a partition of all of X that is finer than the main cascade $\mathcal{C}$, we say that two non-port states x, y are *port-equivalent* iff they are in the same member of $\mathcal{C}$; do not lie in two different eddies, abysses, or trenches; and fulfil $x \rightsquigarrow_X P \Leftrightarrow y \rightsquigarrow_X P$ and $P \rightsquigarrow_X x \Leftrightarrow P \rightsquigarrow_X y$ for all $P \in \mathcal{P}$. Each maximal topologically connected set of port-equivalent states is now called a *rapid*. This ensures that not only U and D are partitioned into ports and rapids but also each individual eddy, abyss, and trench. The ports and rapids together form the *ports-and-rapids partition*, $\mathcal{PR}$, which is finer than $\mathcal{C}$.

A set $H \subseteq X$ is *harbourish* iff it has $x \rightsquigarrow_{X^+} y$ for all x, $y \in H$; is topologically connected, does not intersect two different lakes, eddies, or abysses; and does not intersect two different connected components of $S+G$. A maximal harbourish set is called a *harbour*. Let $\mathcal{H}$ be the system of all harbours. Two non-harbour states x, $y \in X^+$ are *harbour-equivalent* iff they (i) are in the same member of $\{S+G$, L, $U^{(+)}$, W, $D^{(+)}$, E^+, $\Upsilon^+\}$; (ii) do not lie in two different lakes, eddies, or abysses; (iii) do not lie in two different connected components of $S+G$; and (iv) fulfil the equivalences $x \rightsquigarrow_{X^+} H \Leftrightarrow y \rightsquigarrow_{X^+} H$ and $H \rightsquigarrow_{X^+ x} \Leftrightarrow H \rightsquigarrow_{X^+} y$ for all $H \in \mathcal{H}$. Each maximal topologically connected set of harbour-equivalent states is called a *channel* and lies completely in either one port or one rapid (see below for a proof), and hence the resulting *harbours-and-channels partition* of X^+, $\mathcal{HC}$, is finer than $\mathcal{PR}$.

A set $O \subseteq X$ is *dockish* iff it has $x \rightsquigarrow_S y$ for all x, $y \in O$, is topologically connected and does not intersect two different shelters. A maximal dockish set is called a *dock*. Let $\mathcal{O}$ be the system of all docks. Two non-dock states x, $y \in S$ are called *dock-equivalent* iff they belong to the same shelter and $x \rightsquigarrow_S O \Leftrightarrow y \rightsquigarrow_S O$ and $O \rightsquigarrow_S x \Leftrightarrow O \rightsquigarrow_S y$ for all $O \in \mathcal{O}$. Each maximal topologically connected set of dock-equivalent states is called a *fairway* and lies completely in either one harbour or one channel, and hence the resulting *docks-and-fairways partition* of S, $\mathcal{OF}$, is finer than $\mathcal{HC}$.

Proposition 3 (Ports, rapids, harbours, etc.).

1. *Each two ports* [*or harbours or docks*] *are disjoint.*

2. *Each port lies completely in one of* U, D, E, Υ^-, Θ, *no port intersects* Υ^+.

3. *Each harbour* [*or dock*] *lies completely in one port* [*or harbour*].

4. *Each channel* [*or fairway*] *lies completely in one member of* $\mathcal{PR}$ [*or* $\mathcal{HC}$].

5. *These partitions are successive refinements of each other*: $\mathcal{C}$, $\mathcal{PR}$, $\mathcal{HC}$, $\mathcal{OF}$.

6. *If a harbour* H *intersects some of the regions* $S+G$, L, U^+, W, *or* D^+, *it is already completely contained in that region.*

Proof.

1. Assume $y \in A \cap A'$ for two different maximal portish (or harbourish or dockish) sets A, A' and put $B = A + A'$. But then B is itself portish (or harbourish or dockish) because stable reachability is transitive. This contradicts the maximality of A and A'.

2. By Proposition 2, if $x \rightsquigarrow_P y \rightsquigarrow_P x$ then x and y must belong to the same member of $\mathcal{C}$, and hence each port lies completely in one of them.

   To show that a port $P \subseteq \Upsilon$ is already in Υ^-, assume $x \in P \cap \Upsilon^+ \subseteq X^+ \in \mathcal{T}$. We will now construct a contradiction by constructing an admissible trajectory from x that avoids Θ forever. Since $x \rightsquigarrow_X x$ and X^+ is open, there is an open set $A \subseteq X^+$ with $y \rightsquigarrow_X x$ for all $y \in A$. Since τ_x is continuous and A open, we find $t_0 > 0$ with $\tau_x(t) \in A$ for all $t \in [0, t_0]$. Let $y = \tau_x(t_0)$ and pick a $\mu \in \mathcal{M}_y$ that returns arbitrarily closely to x. Let $\mathcal{A}$ be the set of all open $A \subseteq X^+$ with $x \in A$, and choose a $t_A > 0$ with $\mu(t_A) \in A$ for all $A \in \mathcal{A}$ (this requires the axiom of choice, which we will assume here). Let $t_1 = \inf_{A \in \mathcal{A}} \sup_{B \in \mathcal{A}, B \subseteq A} t_B \geqslant 0$. Since $y \in \Upsilon + \Theta$, there is $t' > 0$ with $\mu(t'') \in \Theta$ for all $t'' > t'$, and hence $t_A \leqslant t'$ for all $A \in \mathcal{A}$ and thus $t_1 \leqslant t'$. Next we show that $\mu(t_1) = x$. If $\mu(t_1) = y \neq x$, one can choose $A \in \mathcal{A}$ and $C \in \mathcal{T}$ with $y \in C$ and $A \cap C = \varnothing$ (this is the only point where we need the Hausdorff property). Since μ is continuous, there are $t_l < t_1$ and $t_u > t_1$ with $\mu(t') \in C$ for all $t' \in [t_l, t_u]$. By definition of t_1, there is $A' \in \mathcal{A}$ with $\sup_{B \in \mathcal{A}, B \subseteq A'} t_B \in [t_1, t_u]$. Putting $A'' = A \cap A' \in \mathcal{A}$, we then also have $\sup_{B \in \mathcal{A}, B \subseteq A''} t_B \in [t_1, t_u]$, and hence there is $B \in \mathcal{A}$ with $B \subseteq A'' \subseteq A$ and $t_B \geqslant t_l$ and hence $\mu(t_B) \in C$ by choice of t_l. But $\mu(t_B) \in B \subseteq A$ by choice of t_B. Hence $\mu(t_B) \in A \cap C = \varnothing$, a contradiction. Thus $\mu(t_1) = x$ after all. Finally we concatenate $\tau_x[0, t_0]$ and $\mu[0, t_1]$ infinitely many times and get an admissible trajectory from x that avoids Θ forever.

3. This follows because $\rightsquigarrow_S$ refines $\rightsquigarrow_{X^+}$, which refines $\rightsquigarrow_X$.

4. Since dock equivalence refines harbour equivalence, which refines port equivalence.

5. Follows from points 2–4.

6. This follows directly from the definitions of $S+G$, L, U^+, W, and D^+ by means of $\rightsquigarrow_X$ and $\rightsquigarrow_{X^+}$ and the transitivity of those relations.

Q. E. D.

A4 Remarks

- In general, $A^{\iota\circ}$ may be properly smaller than both the interior $(A^\iota)^\circ$ of the largest invariant subset A^ι of A and

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

the largest invariant subset of A°, $(A^\circ)^\iota$. The three sets can only be shown to be equal under additional smoothness assumptions on τ and $\mu \in \mathcal{M}_x$.

- The set of all states that are stably reachable from x need not be closed or open and need not contain any of the intermediate states that lie on the trajectories $\mu \in \mathcal{M}_x$ used in stable reachability.
- $x \rightsquigarrow_A Y$ does not imply $x \rightsquigarrow y$ for any $y \in Y$, since, after a perturbation, other points in Y may be reachable than before.
- For two points x, y in the same port, harbour, or dock A, one may still not have $x \rightsquigarrow_A y$ since the intermediate states on the trajectories from x to y may not be *stably* reachable from x and thus may not belong to A. In other words, perturbations may still push the system temporarily out of a port, harbour, or dock, but one can then return to the same port, harbour, or dock. For this reason, the directed reachability network is typically acyclic but may contain reachability cycles in pathological situations.
- Any attractor A with the return property (e.g. a stable fixed point or limit cycle, and most strange and chaotic attractors) of the default dynamics lies completely within one port, and hence within one member of $\mathcal{C}$. If $A \subseteq X^+$ then already $A \subseteq S$, and A lies completely within one dock.
- The scope of possible connection topologies that may occur as the reachability network of a managed system contains at least all acyclic finite or countably infinite directed graphs, as can be seen by the following construction: given an acyclic directed graph, one can construct a topologically equivalent network of water bowls which are connected by water tubes leading from a dedicated “drain” at the bottom of the source ball to a common entrance at the top of the target ball. Let water flow into all balls without incoming tubes and out of all outgoing tubes through grilles, determining the default dynamics of a small submarine floating in the water. Then assume the submarine can be propelled strongly enough to move freely inside each ball and to each drain, but not strongly enough to leave the ball through the entrance at the top, against the direction of the water flow. By making parts of the balls and tubes opaque and moving some of the drains from the bottom to the sides of the ball, the construction can be extended to show that also all internally consistent three-level acyclic networks can occur as the three-level network of ports, harbours, and docks.

Appendix B: Further examples

B1 One-dimensional potential function

This simple model shows how almost all of the introduced state space regions (except eddies and dark abysses) may already occur in a one-dimensional system $\dot{x} = -\mathrm{d}f/\mathrm{d}x$ that is defined by a potential function $f(x)$ and already for simple desirable regions such as $X^+ =]0, \infty[$, as depicted in Fig. B1.

Our example has default dynamics along the blue line downwards at a speed proportional to slope, but management is able to move upwards instead on the thin blue lines where the slope is small enough (for $|\mathrm{d}f/\mathrm{d}x| < 3/2$). The chosen undesirable region of $x \leqslant 0$ is indicated in grey. The shelter consists of the two segments just left of point a and it can be stably reached from everywhere properly left of a; hence that whole region constitutes the upstream. The manageable region is the union of shelter, glade, lake, and backwater, and it can be stably reached from everywhere properly left of point b; hence the downstream is the right-open interval from a to b.

That there are no eddies and no dark abysses in this example is typical for systems without any circular flows and with a sufficiently simply shaped X^+.

There are two ports, i.e. the two closed intervals where the default flow is slow: one in the upstream and one in the downstream. Note that the latter is only partially contained in the backwater. One rapid lies to the left of the left port, another between the left port and point a, and these two rapids are port-equivalent since both can reach the left but not the right port. Similarly, the right port is surrounded by two port-equivalent rapids. Finally, there is a singleton rapid consisting only of the point a and a last one formed by point b and all that is to the right of it; from these two port-equivalent rapids, no port can be stably (!) reached.

B2 Bifurcations of a directly manageable flow

If a system passes through a bifurcation, the classification of states by the criteria outlined above will typically change. Let us examine some archetypical cases that can occur in the exemplary case where management can directly affect the flow by changing the default derivative $\dot{x} = F(x)$ of a one-dimensional system by at most one unit, so that the admissible trajectories are those with $\dot{x} \in [F(x) - 1, F(x) + 1]$. (See Sect. 3.6 above for the case where management is via changing a parameter instead.)

Assume $X^+ = \{|x| < \ell\}$ for some $\ell \gg 1$, and the default flow has a *subcritical pitchfork bifurcation*, say $F(x) = x^3 - rx$, where for $r > 0$ the stable fixed point $x_0 = 0$ is surrounded by two unstable ones at $x_\pm = \pm\sqrt{r}$ and becomes unstable itself for $r \leqslant 0$, as depicted by the solid and dotted pale-blue lines in Fig. B2a. Then for $r > 0$, we have a shelter-and-glade situation with a shelter

$S =]-\sqrt{r},\ \sqrt{r}[$ and two glades $G =]-g(r),\ -\sqrt{r}] + [\sqrt{r},\ g(r)[$ where $g(r) > \sqrt{r}$ is the upper solution to the equation $F(g(r)) - 1 = 0$, indicating the limit above which also the extreme management with $\dot{x} = F(x) - 1$ cannot move the system downwards (dashed dark-blue lines). But for $r \leqslant 0$, the shelter disappears and the glades merge and are converted into a backwater $W =]\text{-}g(r),\ g(r)[$. In both cases, this is surrounded by two sunny abysses $\Upsilon^+ =]-\ell,\ -g(r)] + [g(r),\ \ell[$ and two trenches $\Theta =]-\infty,\ \ell] + [\ell,\ \infty[$ (outside the depicted area). One may call this transition a *backwater/glade bifurcation*. As an early warning signal of an imminent breakdown of a shelter in such a backwater/glade bifurcation, one may consider the volume of the shelters Vol(S) in terms of some natural measure on X as a measure of "shelter stability", similar to the concept of basin stability for unmanaged systems without desirable region (Menck et al., 2013; Ji and Kurths, 2014; Schultz et al., 2014; van Kan et al., 2015) and to the recently introduced survivability measure for unmanaged systems with a desirable region (Hellmann et al., 2015).

The port surrounding the unstable fixed point $x = 0$, $P_0 =]-g(r),\ g(r)[$, where $g(r)$ is the solution to $F(g(r)) + 1 = 0$, eventually also splits into three ports P_0 and $P_\pm$, separated by two rapids $R_\pm$; their borders are depicted by the dashed red lines. But this happens only at a larger value of r, namely at $r = 3/\sqrt[3]{4} \approx 1.9$, after which the two unstable fixed points $x_\pm$ can no longer be reached from each other. The corresponding ports-and-rapids network has these arrows: $\neg(P_- \rightsquigarrow_X)\,\neg(R_- \rightsquigarrow_X)\ P_0 \rightsquigarrow_X R_+ \rightsquigarrow_X P_+$. One may call this transition a *port pitchfork bifurcation*.

An interesting case is a *saddle-node bifurcation* such as the one in Fig. B2b, with $F(x) = -r - x^2$ and a critical parameter value $r = 0$ at which the stable and unstable fixed points at $x = \pm\sqrt{-r}$ collide and disappear. First, at the critical point, the shelter caused by the stable fixed point and its glade are transformed into a backwater. Then, somewhat later (at $r = 1$), the maximal value of $\dot{x}$ achievable by management becomes negative and the backwater ceases to exist so that only the sunny abyss remains. One may call this a *glade–backwater–abyss transition*.

If a stable fixed point approaches and eventually enters deeply into the dark region, this may also be called a form of "bifurcation" that causes a similar transition in the classification of states. If $F(x) = -r - x$ and $X^+ = \{x > 0\}$, as in Fig. B2c, then again two changes occur: at $r = 0$, the shelter-and-upstream situation of $r < 0$, with $S =]0,\ \infty[$ and $U^- =]-\infty,\ 0]$, converts into a backwater-and-downstream situation with $W =]0,\ \infty[$ and $D^- =]-\infty,\ 0]$. Then at $r = 1$, this further converts into an abyss-and-trench situation of $r \geqslant 1$ with $\Upsilon^+ =]0,\ \infty[$ and $\Theta =]\infty,\ 0]$. One could thus call this a *shelter–backwater–abyss transition*.

Finally, a transition with three steps is caused if the fixed point passes through a narrower strip of dark, as in Fig. B2d, where again $F(x) = -r - x$ but now $X^+ = \{|x| > 1/4\}$. Here the shelter is again first transformed into a backwater at $r = -1/4$, but then into a lake L when the fixed point leaves the dark again at $r = +1/4$, and even later into a remaining sunny upstream $U^{(+)}$ once the maximally achievable value of $\dot{x}$ at the upper boundary of the dark, i.e. at $x = 1/4$, becomes negative. We suggest to call this a *shelter–backwater–lake–upstream transition*.

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

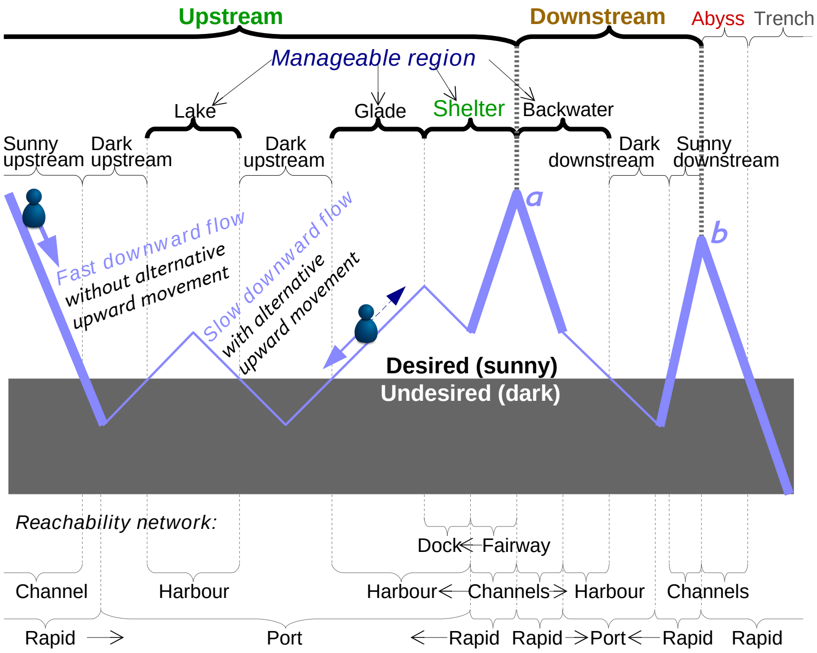


Figure B1. A system moves along the blue line: downward by default (pale-blue arrows), but in some regions management can move it in the opposite direction (dark-blue arrow) in order to avoid the undesired "dark" region. *Shelters, manageable region, upstream, and downstream* (boldface, Sect. 2.2) and other regions from the *main cascade* (top line, Sect. 2.3). Regions from the finer *manageable partition* (below, Sect. 2.4). See Fig. 2 for a systematic summary of these concepts. Bottom: three-level *reachability network* (Sect. 2.5).

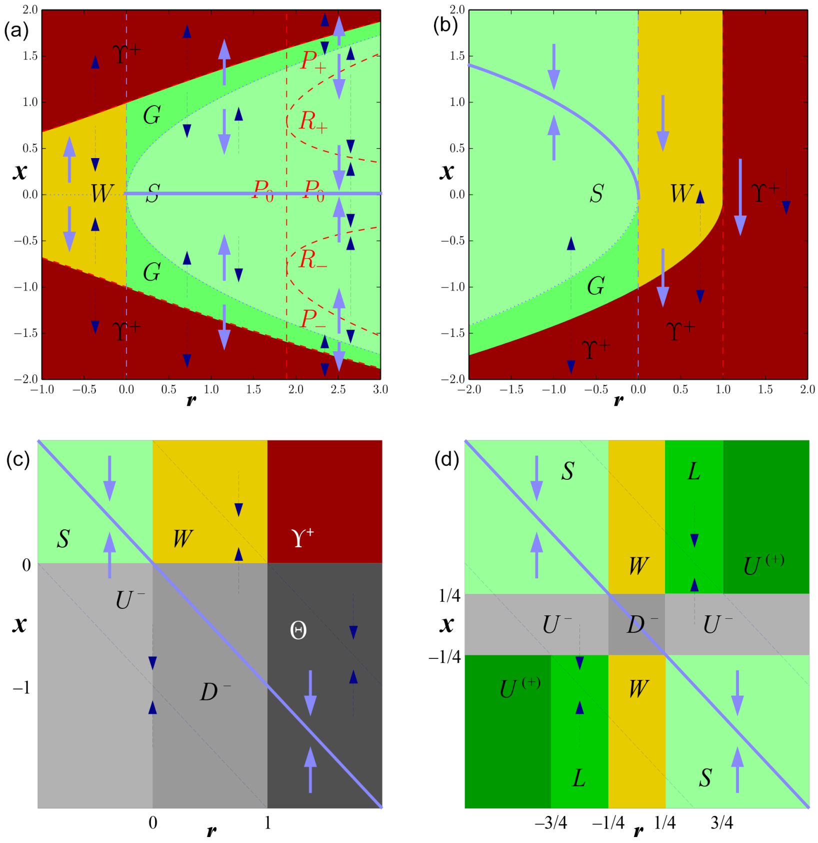
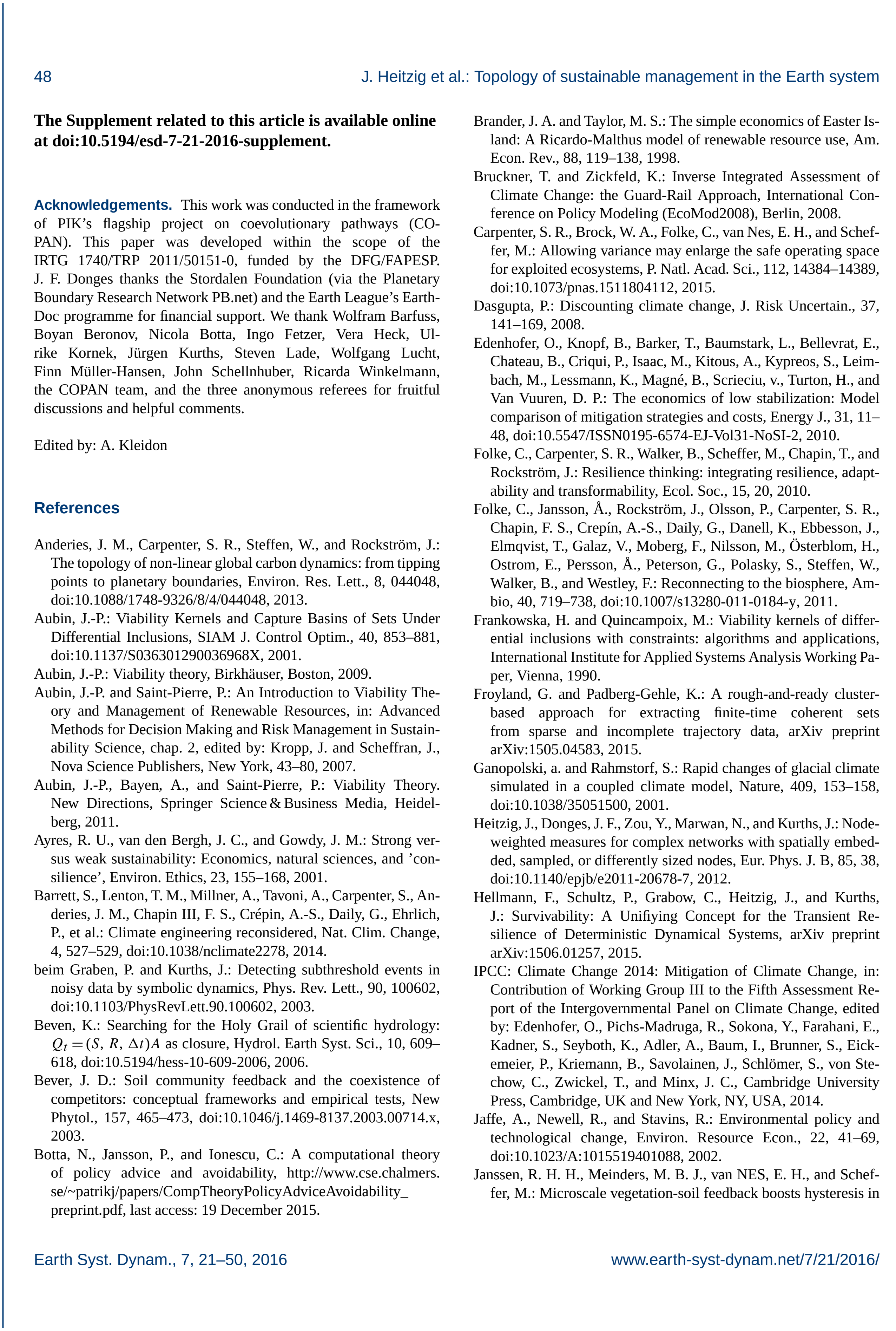


Figure B2. Parameter changes can change the quality of states due to bifurcations. Top-left panel: backwater/glade bifurcation and later port pitchfork bifurcation caused by a subcritical pitchfork bifurcation of the default flow (similar in the supercritical case). Top-right panel: glade–backwater–abyss transition caused by a saddle-node bifurcation, with the second critical value marked in red. Bottom-left panel: shelter–backwater–abyss transition caused by the transition of a stable fixed point into the deep dark. Bottom-right panel: shelter–backwater–lake–upstream transition caused by the transition of a stable fixed point through a dark strip.

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).



The Supplement related to this article is available online at doi:10.5194/esd-7-21-2016-supplement.

Acknowledgements. This work was conducted in the framework of PIK's flagship project on coevolutionary pathways (COPAN). This paper was developed within the scope of the IRTG 1740/TRP 2011/50151-0, funded by the DFG/FAPESP. J. F. Donges thanks the Stordalen Foundation (via the Planetary Boundary Research Network PB.net) and the Earth League's EarthDoc programme for financial support. We thank Wolfram Barfuss, Boyan Beronov, Nicola Botta, Ingo Fetzer, Vera Heck, Ulrike Kornek, Jürgen Kurths, Steven Lade, Wolfgang Lucht, Finn Müller-Hansen, John Schellnhuber, Ricarda Winkelmann, the COPAN team, and the three anonymous referees for fruitful discussions and helpful comments.

Edited by: A. Kleidon

References

Anderies, J. M., Carpenter, S. R., Steffen, W., and Rockström, J.: The topology of non-linear global carbon dynamics: from tipping points to planetary boundaries, Environ. Res. Lett., 8, 044048, doi:10.1088/1748-9326/8/4/044048, 2013.

Aubin, J.-P.: Viability Kernels and Capture Basins of Sets Under Differential Inclusions, SIAM J. Control Optim., 40, 853–881, doi:10.1137/S036301290036968X, 2001.

Aubin, J.-P.: Viability theory, Birkhäuser, Boston, 2009.

Aubin, J.-P. and Saint-Pierre, P.: An Introduction to Viability Theory and Management of Renewable Resources, in: Advanced Methods for Decision Making and Risk Management in Sustainability Science, chap. 2, edited by: Kropp, J. and Scheffran, J., Nova Science Publishers, New York, 43–80, 2007.

Aubin, J.-P., Bayen, A., and Saint-Pierre, P.: Viability Theory. New Directions, Springer Science & Business Media, Heidelberg, 2011.

Ayres, R. U., van den Bergh, J. C., and Gowdy, J. M.: Strong versus weak sustainability: Economics, natural sciences, and 'consilience', Environ. Ethics, 23, 155–168, 2001.

Barrett, S., Lenton, T. M., Millner, A., Tavoni, A., Carpenter, S., Anderies, J. M., Chapin III, F. S., Crépin, A.-S., Daily, G., Ehrlich, P., et al.: Climate engineering reconsidered, Nat. Clim. Change, 4, 527–529, doi:10.1038/nclimate2278, 2014.

beim Graben, P. and Kurths, J.: Detecting subthreshold events in noisy data by symbolic dynamics, Phys. Rev. Lett., 90, 100602, doi:10.1103/PhysRevLett.90.100602, 2003.

Beven, K.: Searching for the Holy Grail of scientific hydrology: $Q_t = (S, R, \Delta t)A$ as closure, Hydrol. Earth Syst. Sci., 10, 609–618, doi:10.5194/hess-10-609-2006, 2006.

Bever, J. D.: Soil community feedback and the coexistence of competitors: conceptual frameworks and empirical tests, New Phytol., 157, 465–473, doi:10.1046/j.1469-8137.2003.00714.x, 2003.

Botta, N., Jansson, P., and Ionescu, C.: A computational theory of policy advice and avoidability, http://www.cse.chalmers.se/~patrikj/papers/CompTheoryPolicyAdviceAvoidability_preprint.pdf, last access: 19 December 2015.

Brander, J. A. and Taylor, M. S.: The simple economics of Easter Island: A Ricardo-Malthus model of renewable resource use, Am. Econ. Rev., 88, 119–138, 1998.

Bruckner, T. and Zickfeld, K.: Inverse Integrated Assessment of Climate Change: the Guard-Rail Approach, International Conference on Policy Modeling (EcoMod2008), Berlin, 2008.

Carpenter, S. R., Brock, W. A., Folke, C., van Nes, E. H., and Scheffer, M.: Allowing variance may enlarge the safe operating space for exploited ecosystems, P. Natl. Acad. Sci., 112, 14384–14389, doi:10.1073/pnas.1511804112, 2015.

Dasgupta, P.: Discounting climate change, J. Risk Uncertain., 37, 141–169, 2008.

Edenhofer, O., Knopf, B., Barker, T., Baumstark, L., Bellevrat, E., Chateau, B., Criqui, P., Isaac, M., Kitous, A., Kypreos, S., Leimbach, M., Lessmann, K., Magné, B., Scrieciu, v., Turton, H., and Van Vuuren, D. P.: The economics of low stabilization: Model comparison of mitigation strategies and costs, Energy J., 31, 11–48, doi:10.5547/ISSN0195-6574-EJ-Vol31-NoSI-2, 2010.

Folke, C., Carpenter, S. R., Walker, B., Scheffer, M., Chapin, T., and Rockström, J.: Resilience thinking: integrating resilience, adaptability and transformability, Ecol. Soc., 15, 20, 2010.

Folke, C., Jansson, Å., Rockström, J., Olsson, P., Carpenter, S. R., Chapin, F. S., Crepín, A.-S., Daily, G., Danell, K., Ebbesson, J., Elmqvist, T., Galaz, V., Moberg, F., Nilsson, M., Österblom, H., Ostrom, E., Persson, Å., Peterson, G., Polasky, S., Steffen, W., Walker, B., and Westley, F.: Reconnecting to the biosphere, Ambio, 40, 719–738, doi:10.1007/s13280-011-0184-y, 2011.

Frankowska, H. and Quincampoix, M.: Viability kernels of differential inclusions with constraints: algorithms and applications, International Institute for Applied Systems Analysis Working Paper, Vienna, 1990.

Froyland, G. and Padberg-Gehle, K.: A rough-and-ready cluster-based approach for extracting finite-time coherent sets from sparse and incomplete trajectory data, arXiv preprint arXiv:1505.04583, 2015.

Ganopolski, a. and Rahmstorf, S.: Rapid changes of glacial climate simulated in a coupled climate model, Nature, 409, 153–158, doi:10.1038/35051500, 2001.

Heitzig, J., Donges, J. F., Zou, Y., Marwan, N., and Kurths, J.: Node-weighted measures for complex networks with spatially embedded, sampled, or differently sized nodes, Eur. Phys. J. B, 85, 38, doi:10.1140/epjb/e2011-20678-7, 2012.

Hellmann, F., Schultz, P., Grabow, C., Heitzig, J., and Kurths, J.: Survivability: A Unifiying Concept for the Transient Resilience of Deterministic Dynamical Systems, arXiv preprint arXiv:1506.01257, 2015.

IPCC: Climate Change 2014: Mitigation of Climate Change, in: Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Farahani, E., Kadner, S., Seyboth, K., Adler, A., Baum, I., Brunner, S., Eickemeier, P., Kriemann, B., Savolainen, J., Schlömer, S., von Stechow, C., Zwickel, T., and Minx, J. C., Cambridge University Press, Cambridge, UK and New York, NY, USA, 2014.

Jaffe, A., Newell, R., and Stavins, R.: Environmental policy and technological change, Environ. Resource Econ., 22, 41–69, doi:10.1023/A:1015519401088, 2002.

Janssen, R. H. H., Meinders, M. B. J., van NES, E. H., and Scheffer, M.: Microscale vegetation-soil feedback boosts hysteresis in

Reproduced from: J. Heitzig et al., Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dyn., vol. 7, no. 1, pp. 21-50, 2016, doi: 10.5194/esd-7-21-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

a regional vegetation–climate system, Global Change Biol., 14, 1104–1112, doi:10.1111/j.1365-2486.2008.01540.x, 2008.

Ji, P. and Kurths, J.: Basin stability of the Kuramoto-like model in small networks, Eur. Phys. J. Spec. Top., 223, 2483–2491, doi:10.1140/epjst/e2014-02213-0, 2014.

Kalkuhl, M., Edenhofer, O., and Lessmann, K.: Learning or lock-in: Optimal technology policies to support mitigation, Resour. Energy Econ., 34, 1–23, doi:10.1016/j.reseneeco.2011.08.001, 2012.

Keller, K., Hall, M., Kim, S. R., Bradford, D. F., and Oppenheimer, M.: Avoiding dangerous anthropogenic interference with the climate system, Climatic Change, 73, 227–238, doi:10.1007/s10584-005-0426-8, 2005.

Kleidon, A. and Renner, M.: A simple explanation for the sensitivity of the hydrologic cycle to surface temperature and solar radiation and its implications for global climate change, Earth Syst. Dynam., 4, 455–465, doi:10.5194/esd-4-455-2013, 2013.

Kleidon, A., Kravitz, B., and Renner, M.: The hydrological sensitivity to global warming and solar geoengineering derived from thermodynamic constraints, Geophys. Res. Lett., 42, 138–144, doi:10.1002/2014GL062589, 2015.

Kourtev, P. S., Ehrenfeld, J. G., and Häggblom, M.: Exotic Plant Species Alter the Microbial Community Structure and Function in the Soil, Ecology, 83, 3152–3166, doi:10.1890/0012-9658(2002)083[3152:EPSATM]2.0.CO;2, 2002.

Kreps, D. M.: A Representation Theorem for "Preference for Flexibility", Econometrica, 47, 565–577, doi:10.2307/1910406, 1979.

Kuipers, B.: Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge, MIT Press, Cambridge, MA, 1994.

Kulmatiski, A., Heavilin, J., and Beard, K. H.: Testing predictions of a three-species plant-soil feedback model, J. Ecol., 99, 542–550, doi:10.1111/j.1365-2745.2010.01784.x, 2011.

Lade, S. J., Tavoni, A., Levin, S. A., and Schlüter, M.: Regime shifts in a social-ecological system, Theor. Ecol., 6, 359–372, doi:10.1007/s12080-013-0187-3, 2013.

Lade, S. J., Niiranen, S., Hentati-Sundberg, J., Blenckner, T., Boonstra, W. J., Orach, K., Quaas, M. F., Österblom, H., and Schlüter, M.: An empirical model of the Baltic Sea reveals the importance of social dynamics for ecological regime shifts, P. Natl. Acad. Sci., 112, 11120–11125, doi:10.1073/pnas.1504954112, 2015a.

Lade, S. J., Niiranen, S., and Schlüter, M.: Generalized modeling of empirical social-ecological systems, arXiv preprint arXiv:1503.02846, 2015b.

Lenton, T. M. and Vaughan, N. E.: The radiative forcing potential of different climate geoengineering options, Atmos. Chem. Phys., 9, 5539–5561, doi:10.5194/acp-9-5539-2009, 2009.

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J.: Tipping elements in the Earth's climate system, P. Natl. Acad. Sci. USA, 105, 1786–1793, doi:10.1073/pnas.0705414105, 2008.

Levine, J. M., Pachepsky, E., Kendall, B. E., Yelenik, S. G., and Lambers, J. H. R.: Plant-soil feedbacks and invasive spread, Ecol. Lett., 9, 1005–1014, doi:10.1111/j.1461-0248.2006.00949.x, 2006.

Martin, S.: The cost of restoration as a way of defining resilience: a viability approach applied to a model of lake eutrophication, Ecol. Soc., 9, 8, 2004.

Menck, P. J., Heitzig, J., Marwan, N., and Kurths, J.: How basin stability complements the linear-stability paradigm, Nat. Phys., 9, 89–92, doi:10.1038/nphys2516, 2013.

Mitra, C., Kurths, J., and Donner, R. V.: An integrative quantifier of multistability in complex systems based on ecological resilience, Scient. Rep., 5, 1–12, doi:10.1038/srep16196, 2015.

Nagy, B., Farmer, J. D., Bui, Q. M., and Trancik, J. E.: Statistical Basis for Predicting Technological Progress, PLoS ONE, 8, 1–7, doi:10.1371/journal.pone.0052669, 2013.

Nicolis, C.: Long-term climatic variability and chaotic dynamics, Tellus A, 39, 1–9, doi:10.3402/tellusa.v39i1.11734, 1987.

Nocke, T., Buschmann, S., Donges, J. F., Marwan, N., Schulz, H.-J., and Tominski, C.: Review: visual analytics of climate networks, Nonlin. Processes Geophys., 22, 545–570, doi:10.5194/npg-22-545-2015, 2015.

Padberg, K., Thiere, B., Preis, R., and Dellnitz, M.: Local expansion concepts for detecting transport barriers in dynamical systems, Commun. Nonlin. Sci. Numer. Simul., 14, 4176–4190, doi:10.1016/j.cnsns.2009.03.018, 2009.

Petschel-Held, G., Schellnhuber, H.-J., Bruckner, T., Toth, F. L., and Hasselmann, K.: The tolerable windows approach: theoretical and methodological foundations, Climatic Change, 41, 303–331, doi:10.1023/A:1019080704864, 1999.

Poon, G. T.: The influence of soil feedback and plant traits on competition between an invasive plant and co-occurring native and exotic species, Master's thesis, University of Guelph, Guelph, 2011.

Rahmstorf, S., Crucifix, M., Ganopolski, a., Goosse, H., Kamenkovich, I., Knutti, R., Lohmann, G., Marsh, R., Mysak, L., Wang, Z., and a.J. Weaver: Thermohaline circulation hysteresis: a model intercomparison, Geophys. Res. Lett., 32, 1–5, doi:10.1029/2005GL023655, 2005.

Raworth, K.: A Safe and Just Space For Humanity: Can we live within the Doughnut?, Oxfam Policy and Practice: Climate Change and Resilience, 8, 1–26, doi:10.5822/978-1-61091-458-1, 2012.

Read, D. B., Bengough, A. G., Gregory, P. J., Crawford, J. W., Robinson, D., Scrimgeour, C. M., Young, I. M., Zhang, K., and Zhang, X.: Plant Roots Release Phospholipid Surfactants That Modify the Physical and Chemical Properties of Soil, New Phytol.t, 157, 315–326, doi:10.1046/j.1469-8137.2003.00665.x, 2003.

Rockström, J., Steffen, W., Noone, K., and Persson, A.: Planetary boundaries: exploring the safe operating space for humanity, Ecol. Soc., 14, 32, 2009a.

Rockström, J., Steffen, W., Noone, K., Persson, A., Chapin, F. S., Lambin, E. F., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R. W., Fabry, V. J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., and Foley, J. A.: A safe operating space for humanity, Nature, 461, 472–475, doi:10.1038/461472a, 2009b.

Rougé, C., Mathias, J. D., and Deffuant, G.: Extending the viability theory framework of resilience to uncertain dynamics, and application to lake eutrophication, Ecol. Indicat., 29, 420–433, doi:10.1016/j.ecolind.2012.12.032, 2013.

Saltzman, B., Sutera, A., and Hansen, A. R.: A Possible Marine Mechanism for Internally Generated Long-Period Climate Cycles, J. Atmos. Sci., 39, 2634–2637, doi:10.1175/1520-0469(1982)039<2634:APMMFI>2.0.CO;2, 1982.

Scheffer, M., Barrett, S., Carpenter, S. R., Folke, C., Green, A. J., Holmgren, M., Hughes, T. P., Kosten, S., van de Leemput, I. A., Nepstad, D. C., van Nes, E. H., Peeters, E. T. H. M., and Walker, B.: Creating a safe operating space for iconic ecosystems, Science, 347, 1317–1319, doi:10.1126/science.aaa3769, 2015.

Schellnhuber, H. J.: Discourse: Earth System Analysis – The Scope of the Challenge, in: Earth System Analysis: Integrating Science for Sustainability, chap. 1, edited by: Schellnhuber, H. J. and Wenzel, V., Springer, Berlin, Heidelberg, 3–195, doi:10.1007/978-3-642-52354-0_1, 1998.

Schellnhuber, H.-J.: 'Earth system' analysis and the second Copernican revolution, Nature, 402, C19–C23, 1999.

Schellnhuber, H. J.: Tipping elements in the Earth System, P. Natl. Acad. Sci. USA, 106, 20561, doi:10.1073/pnas.0911106106, 2009.

Schultz, P., Heitzig, J., and Kurths, J.: Detours around basin stability in power networks, New J. Phys., 16, 125001, doi:10.1088/1367-2630/16/12/125001, 2014.

Ser-Giacomi, E., Rossi, V., López, C., and Hernández-García, E.: Flow networks: A characterization of geophysical fluid transport, Chaos, 25, 036404, doi:10.1063/1.4908231, 2015.

Singh, R., Reed, P. M., and Keller, K.: Many-objective robust decision making for managing an ecosystem with a deeply uncertain threshold response, Ecol. Soc., 20, 1–32, 2015.

Sontag, E. D.: Mathematical control theory: Deterministic Finite Dimensional Systems, 2nd Edn., Springer, New York, 1998.

Steffen, W., Richardson, K., Rockström, J., Cornell, S., Fetzer, I., Bennett, E., Biggs, R., Carpenter, S. R., de Wit, C. a., Folke, C., Mace, G., Persson, L. M., Veerabhadran, R., Reyers, B., and Sörlin, S.: Planetary Boundaries: Guiding human development on a changing planet, Science, 347, 1–17, doi:10.1126/science.1259855, 2015.

Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M.: Climate Change 2013: The Physical Science Basis, Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, UK and New York, NY, USA, doi:10.1017/CBO9781107415324, 2013.

Stommel, H.: Thermohaline Convection with Two Stable Regimes of Flow, Tellus, 13, 224–230, doi:10.1111/j.2153-3490.1961.tb00079.x, 1961.

van Kan, A., Jegminat, J., Donges, J. F., and Kurths, J.: Constrained basin stability for studying transient dynamics in complex systems, in review, 2015.

Vaughan, N. E. and Lenton, T. M.: A review of climate geoengineering proposals, Climatic Change, 109, 745–790, doi:10.1007/s10584-011-0027-7, 2011.

Chaos ARTICLE scitation.org/journal/cha

Deep reinforcement learning in World-Earth system models to discover sustainable management strategies

Cite as: Chaos **29**, 123122 (2019); doi: 10.1063/1.5124673
Submitted: 15 August 2019 · Accepted: 20 November 2019 ·
Published Online: 17 December 2019

View Online Export Citation CrossMark

Felix M. Strnad,[1,2,a)] Wolfram Barfuss,[3,4] Jonathan F. Donges,[3,5] and Jobst Heitzig[1]

AFFILIATIONS

[1] FutureLab on Game Theory and Networks of Interacting Agents, Research Department 4: Complexity Science, Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany
[2] Department of Physics, University of Göttingen, 37077 Göttingen, Germany
[3] FutureLab on Earth Resilience in the Anthropocene, Research Department 1: Earth System Analysis, Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany
[4] Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany
[5] Stockholm Resilience Center, Stockholm University, 104 05 Stockholm, Sweden

Note: This paper is part of the Focus Issue, "When Machine Learning Meets Complex Systems: Networks, Chaos and Nonlinear Dynamics."

[a)] **Electronic mail:** strnad@pik-potsdam.de

ABSTRACT

Increasingly complex nonlinear World-Earth system models are used for describing the dynamics of the biophysical Earth system and the socioeconomic and sociocultural World of human societies and their interactions. Identifying pathways toward a sustainable future in these models for informing policymakers and the wider public, e.g., pathways leading to robust mitigation of dangerous anthropogenic climate change, is a challenging and widely investigated task in the field of climate research and broader Earth system science. This problem is particularly difficult when constraints on avoiding transgressions of planetary boundaries and social foundations need to be taken into account. In this work, we propose to combine recently developed machine learning techniques, namely, deep reinforcement learning (DRL), with classical analysis of trajectories in the World-Earth system. Based on the concept of the agent-environment interface, we develop an agent that is generally able to act and learn in variable manageable environment models of the Earth system. We demonstrate the potential of our framework by applying DRL algorithms to two stylized World-Earth system models. Conceptually, we explore thereby the feasibility of finding novel global governance policies leading into a safe and just operating space constrained by certain planetary and socioeconomic boundaries. The artificially intelligent agent learns that the timing of a specific mix of taxing carbon emissions and subsidies on renewables is of crucial relevance for finding World-Earth system trajectories that are sustainable in the long term.

We propose a framework for using deep reinforcement learning (DRL) as an approach to extend the field of Earth system analysis by a new method. We build our framework upon the agent-environment interface concept. The agent can apply management options to models of the Earth system as the environment of interest and learn by rewards provided by the environment. We train our agent with a deep Q-neural network extended by current state-of-the-art algorithms. We find that the agent is able to learn novel, previously undiscovered policies that navigate the system into sustainable regions in two exemplary conceptual models of the World-Earth system.

I. INTRODUCTION

Efforts invested in identifying pathways toward global sustainability need to account for critical feedback loops between the socioeconomic and sociocultural World and the biophysical Earth system.[1,2] These pathways may require novel, yet undiscovered, multilevel policies, from the local to the global scale, for the governance of this coupled World-Earth system leading toward a safe and just operating space.[3,4] Striving for a safe and just operating space, policymakers of the United Nations agreed on global political cooperation for a sustainable future at the resolution of the 17 Sustainable Development Goals (SDG)[5] and the adoption of the Paris Agreement on

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

Climate Change.[6] The safe and just operating space is based on a set of biophysical planetary boundaries (defined on dimensions such as climate change or biosphere integrity loss) as they are formulated by Rockström *et al.* in Refs. 3, 4, 7, and 8, extended by social foundations (e.g., poverty alleviation) by Raworth.[9] If respected together, staying within these boundaries is seen as a prerequisite to ensuring sustainable human development. The field of Earth system modeling develops computer models to show possible pathways toward a sustainable future. However, the identification and characterization of concrete trajectories within the planetary boundaries and above social foundations remains a problem requiring ongoing research efforts.[10,11]

In this paper, we consider this problem on a globally aggregated level assuming the following basic structure: An abstract single decision-maker interacts with a dynamical, in most cases, nonlinear environment to find sustainable trajectories within certain boundaries. The field of *Integrated Assessment Modeling* (IAM) addresses this issue via optimizing a social welfare function in order to estimate the design of sustainable management strategies.[12] IAM models integrate data and knowledge from established climate models.[13,14] To identify pathways in IAM, numerical solvers such as GAMS[15] are frequently used. However, these IAM models are highly dependent on the choice of the target function of the optimization. In many cases, this choice may not be obvious and depends on the IAM developers.[16]

As another approach, *optimal control theory* (OCT) can be used to solve problems where dynamical systems are supposed to stay within certain constraints. In these systems, OCT tries to find an optimal choice for some control variable by optimizing a specific objective function.[17] Applied to Earth system models, the focus has been set on the design of climate regulators and their impact on climate modification.[18,19] *Viability theory* (VT) as a subfield of OCT can be stated as an example. In this field, such problems of identifying trajectories are typically addressed by methods that rely on a discretization of the state space, followed by the application of local linear approximations.[20] It is, however, not well applicable in systems with more than just a small number of variables due to the curse of dimensionality.[21]

The use of *reinforcement learning* (RL)[22] can also be considered as a possible approach for intelligent decision-making within World-Earth system models.[23] It is designed for finding optimal policy strategies as well. However, in contrast to the previously presented approaches, RL does not detect solutions based on numerically solving an optimization problem, but by a dynamic search process via exploration and exploitation of past experiences, guided by a reward function. However, tabular methods, which are mainly used for classical RL solutions, cannot be straightforwardly applied to the systems of interest here, due to the continuous state spaces that we mostly find in World-Earth system models.

The common point of all the presented methods outlined above is that they reach their limits as the complexity of the environments increases. However, *deep reinforcement learning* (DRL)[24] algorithms have been shown to detect solutions in other highly complex environments surprisingly well.[24,25] In this paper, we propose using DRL as a novel approach for Earth system analysis. Even though first successful reinforcement learning experiments by using neural networks as nonlinear function approximators were reported already in 1995,[26] the breakthrough of DRL was achieved only in 2013.[24,25] Since then, DRL algorithms have become increasingly popular in the field of artificial intelligence.[27,28] The key to the success of this approach lies in the combination of Q-learning,[29] neural networks,[30] and experience replay,[31] which has been shown to learn policies up to a superhuman performance in a variety of different environments.[24,25] Often DRL applications come up with unexpected and novel solutions.[32,33] Many extensions have been proposed addressing both speed and efficiency.[34] Due to its general applicability to various environments, DRL is used in a wide range of different fields, e.g., resources management in computer clusters,[35] optimization of chemical reactions,[36] playing abstract strategy games like chess and Go,[32,33] autonomous driving,[37] and, in particular, robotics.[38–41]

Due to the wide applicability of DRL, we propose a framework that uses DRL as a tool that is both robust and easy to use at the same time to identify and classify trajectories in Earth system models effectively. As a proof of concept, we use our DRL framework within various stylized World-Earth system models.[2,42] These models are designed to investigate the coevolutionary dynamics of humans and nature in the Anthropocene. Some first applications of reinforcement learning methods within resource use models have been carried out,[43–45] but as far as we know, there are no approaches yet applying DRL to Earth system models. We believe this approach will open so far unused possibilities to discover so far unknown management strategies that keep the Earth system within planetary boundaries, while, at the same time, respecting social foundations of the world's societies. Recently, various ways of how to tackle problems related to anthropogenic climate change by using machine learning techniques have been outlined.[46] Our work proposes a novel strand to this list.

II. METHODS

This work uses the agent-environment interface[22] as a formal mathematical framework which allows for making a fruitful connection between reinforcement learning and the modeling of social-ecological systems, as it was, e.g., proposed by Barfuss, Donges, and Kurths.[47] In the case of a single agent as studied here, RL problems are based on the concept of Markov decision processes (MDPs).[22] Therefore, we will provide a brief introduction to MDPs, followed by a description of how we included the learning process by using neural networks. We will further give a short overview of possible extensions and conclude this section by outlining how we translate Earth system models into the formal framework of an MDP.

A. Markov decision processes

RL is designed for problems where an agent observing the environment output consisting of a state and a scalar reward signal is acting upon this observation.[22] Formally, this interaction is described by a so-called Markov decision process (MDP).[48] At each step t, the environment is in a certain state $s_t \in \mathcal{S}$, where $\mathcal{S}$ describes the set of all possible states. The agent chooses an action a_t from a given finite action set $a_t \in \mathcal{A}$. Environmental dynamics are now determined by the transition probability $T(s'|a,s) = P(s_{t+1} = s'|s_t = s, a = a_t)$, which does not depend on t explicitly. When for a given action a, the environment transits from state s to s', the agent receives an immediate numerical value r_t, called the *reward*, that generally depends on the state s and action a. The tuple (s, a, r, s') defines the MDP. The

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

agent chooses its action according to its behavior policy π which maps state s to a probability distribution over all actions $a \in \mathcal{A}$, expressed as $\pi(s, a) = P(a|s)$.

B. Deep reinforcement learning

Every decision the agent takes is followed by a reward it gets from the environment. In all types of RL algorithms, the goal of an agent is to maximize its exponentially discounted sum of future rewards,[22] called the gain $G_t = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$, where the discount factor $\gamma \in [0, 1]$ expresses how much the agent cares for future rewards. This lets us define a state-action value function Q_π quantifying the value of a state s, given that the agent applies action a, as the expected return, following a given policy π, $Q_\pi(s) = \mathbb{E}_\pi[G_t|s_t = s, a_t = a]$. The average $\mathbb{E}_\pi$ can be understood as the sum over all actions for a policy π times the sum over all possible state transitions to s'. Inserting the gain G_t yields the *Bellman equation*,[49]

$$Q_\pi(s, a) = \mathbb{E}_\pi\left[r_t + \gamma Q_\pi(s_{t+1}, a_{t+1})|s_t = s, a_t = a\right]. \tag{1}$$

The best possible solution of an MDP is the optimal state-action value function $Q^*(s, a)$, which is the maximum state-action value function over all policies,

$$Q^*(s, a) = \max_\pi Q_\pi(s, a). \tag{2}$$

The problem of maximizing the expected discounted reward sum G_t is transformed to find the optimal state-action value function Q^*. The optimal value function allows the following consideration. If for all possible actions a' for the next time step $s' = s_{t+1}$ the value of $Q^*(s', a')$ was known, then the optimal strategy would be to choose that $a' \in \mathcal{A}$, resulting in the highest value of $Q^*(s', a')$. This identity is known as the *Bellman Optimality Equation*,[22]

$$Q^*(s, a) = \mathbb{E}_T\left[r + \gamma \max_{a' \in \mathcal{A}} Q^*(s', a')\right]. \tag{3}$$

$\mathbb{E}_T$ averages over all possible state transitions, given by the transition probability T. The task is now to find a way to estimate the optimal action value function $Q^*(s, a)$. Estimating the state-action value function by performing rollouts on the environment are called *model-free*. Mnih *et al.*[24,25] address this issue with the combination of Q-learning, deep neural networks, and experience replay successfully, called *deep Q-learning* (DQL). Briefly, we will provide an overview of their approach.

a. Q-learning. Q-learning is a specific type of RL which converges to the optimal solution. In Q-learning, we use the function $Q(s, a)$ representing the state-action value when performing action a in state s. The temporal difference error of expected value $Q(s, a)$ and experienced value $r + Q(s', a')$ is used to estimate the current value of the state.[22] It is used to incrementally estimate Q-values for actions, based on an iteratively updated Q-value function,[29]

$$Q_{i+1}(s_t, a) = r_t + \gamma \max_{a' \in \mathcal{A}} Q_i(s_{t+1}, a'). \tag{4}$$

Action selection when acting in the environment is usually made with an ϵ-greedy policy, i.e., with probability $\epsilon \in [0, 1]$ the action $\operatorname{argmax}_a Q(s, a)$ is used, and with probability $1 - \epsilon$ a random action is used. Here, the parameter ϵ regulates this exploration-exploitation trade-off. Q-learning is an *off-policy* algorithm, i.e., to estimate the current state-action value the agent uses the maximum state-action value of the next state, regardless of which action is actually chosen there. Still, one can prove that, for $i \to \infty$, this algorithm will converge to the optimal action value function $Q(s, a) \to Q^*(s, a)$.[22]

b. Deep Q-networks. In practice, this convergence is only applicable in state spaces with a small number of states. However, continuous state spaces make it impossible to learn state-action pairs independently.[22] Using multilayered neural networks as function approximators, $Q(s, a, \theta) \approx Q^*(s, a)$, called deep Q-networks (DQN), is a possibility to overcome this issue.[24] As target function Y_t, one can use different RL variants.[22] Here, the Q-learning update from Eq. (4) is adjusted by setting $Y_t(s_t, a_t, \theta) = r_t + \gamma \max_{a' \in \mathcal{A}} Q(s_{t+1}, a,' \theta)$. The parameters (i.e., the weights) θ_i of the neural network are optimized by gradient descent to minimize the loss $\mathcal{L}_i(\theta_i)$ at iteration i between the target and the current Q value via

$$\mathcal{L}_i(\theta_i) = \left(Y_t(\theta_i^-) - Q(s_t, a_t, \theta_i)\right)^2, \tag{5}$$

$$\nabla_{\theta_i}\mathcal{L}_i = \left(Y_t(\theta_i^-) - Q(s_t, a_t, \theta_i)\right)\nabla_{\theta_i}Q(s_t, a, \theta_i), \tag{6}$$

$$\theta_{i+1} = \theta_i + \alpha\nabla_{\theta_i}\mathcal{L}_i. \tag{7}$$

The parameter α describes the learning rate of the network. To account for a more stable learning, a second network with parameters θ_t^- is used. This network is a copy of the first one but is frozen in time for τ_{target} iteration steps. It is used as the target function $Y_t(s, a, \theta^-)$ in Eq. (5). The fixed Q-values of $Y_t(s, a, \theta^-)$ make it possible that the optimization process converges to a stable target.[24] The target network is updated every τ_{target} iteration steps by copying the parameters from the current network: $\theta^- \leftarrow \theta$.

c. Experience replay. Instead of learning from state-action pairs as they occur during simulation, updates for the state-action value function $Q(s, a, \theta)$ are applied on samples (called Mini-Batches) randomly drawn from a replay memory—typically a large table of stored observations that are collected during the training process. This separates the learning process itself from gaining experience,[31] which breaks the similarity of subsequent training samples and leads consequently to more stable learning.[24]

C. Extensions to DQN

After the convincing performance of the DQN network presented by Mnih *et al.*,[24] the algorithm has been further developed and significant improvements regarding stability and learning speed could be achieved. By using double Q-learning,[50] harmful overestimation of the Q values[51] can be reduced. With the introduction of dueling network architectures,[52] the value of a state and the advantage of taking a certain action at that state could be decoupled. Furthermore, the distributional DQL algorithm by Bellemare, Dabney, and Munos[53] addresses the issue that the value of future rewards is restricted to the expected return (i.e., to the Q function) and replaces it with a distribution of Q-values per action.

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

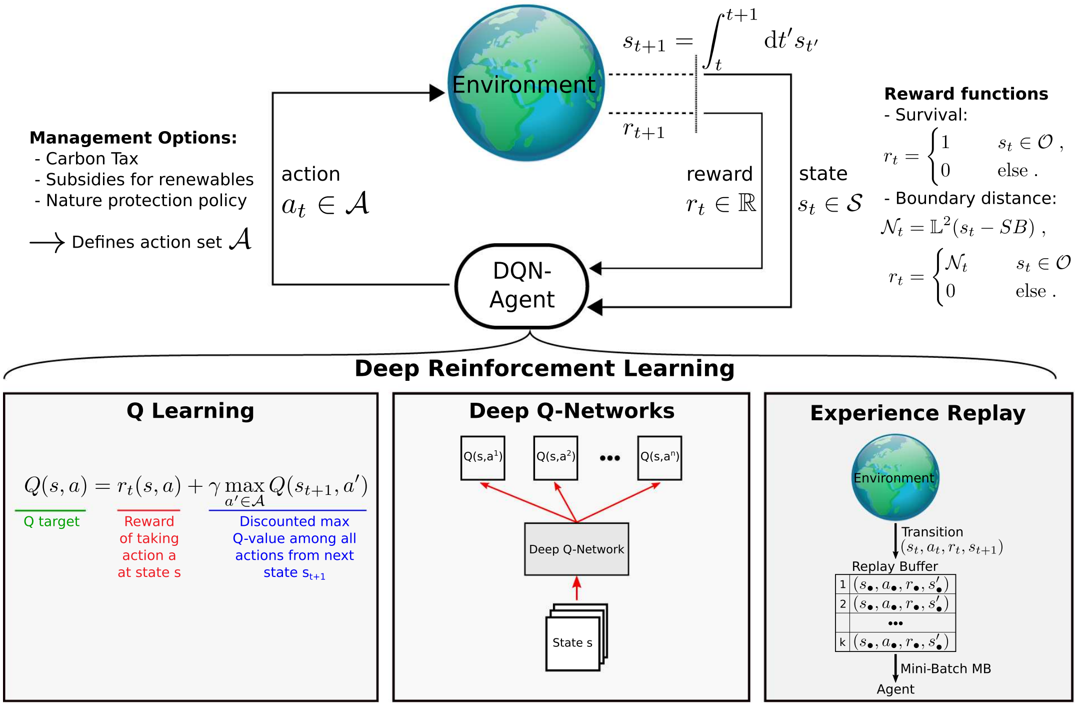

Environment
$s_{t+1} = \int_t^{t+1} \mathrm{d}t' s_{t'}$
r_{t+1}
Management Options:
- Carbon Tax
- Subsidies for renewables
- Nature protection policy
⟶ Defines action set $\mathcal{A}$
action
$a_t \in \mathcal{A}$
reward
$r_t \in \mathbb{R}$
state
$s_t \in \mathcal{S}$
Reward functions
- Survival:
$r_t = \begin{cases} 1 & s_t \in \mathcal{O}\ , \\ 0 & \text{else}\ . \end{cases}$
- Boundary distance:
$\mathcal{N}_t = \mathbb{L}^2(s_t - SB)\ ,$
$r_t = \begin{cases} \mathcal{N}_t & s_t \in \mathcal{O}\ , \\ 0 & \text{else}\ . \end{cases}$
DQN-Agent
Deep Reinforcement Learning
Q Learning
$Q(s,a) = r_t(s,a) + \gamma \max_{a' \in \mathcal{A}} Q(s_{t+1}, a')$
Q target
Reward of taking action a at state s
Discounted max Q-value among all actions from next state s_{t+1}
Deep Q-Networks
Q(s,a¹)
Q(s,a²)
•••
Q(s,aⁿ)
Deep Q-Network
State s
Experience Replay
Environment
Transition
(s_t, a_t, r_t, s_{t+1})
Replay Buffer
1 $(s_\bullet, a_\bullet, r_\bullet, s'_\bullet)$
2 $(s_\bullet, a_\bullet, r_\bullet, s'_\bullet)$
•••
k $(s_\bullet, a_\bullet, r_\bullet, s'_\bullet)$
Mini-Batch MB
Agent

FIG. 1. Using the agent-environment interface[22] for analyzing World-Earth models via deep reinforcement learning (DRL) techniques. The environment is in a certain state s_t, based on that state the agent chooses an action a_t. The environment responds with a next state s_{t+1} and a reward r_{t+1}. The dynamics of the environment for every time step *dt* are numerically integrated. We interpret the action set $\mathcal{A}$ as a set of management options for the Earth system and propose different reward functions. $\mathcal{O}$ describes the set of states that are within the planetary boundaries (PBs). The learning agent is implemented to use DRL[24,25] using Q-learning[29] combined with deep neural networks[30] and experience replay[31] to choose at every step one action from an action set which in our case represents governance management options. In the Q-learning box, the representation of the target function is depicted. To visualize deep Q-networks' functionality, we show a scheme for the function approximator via a deep neural network. In the experience replay box, the dot as the index of the observations in the replay buffer of size *k* represents an arbitrary time point.

After having presented possible improvements to *Q*-learning and the network architectures, one can also optimize the way which experiences are used for learning. When treating all samples the same, we are not considering that possibly we can learn more from some transitions $s \rightarrow s'$ than from others. Prioritized experience replay (PER), developed by Schaul *et al.*,[54] is one strategy that tries to account for this issue by changing the sampling distribution. The basic idea is to use the absolute temporal difference error to prioritize important transitions. However, when PER is introduced, there is obviously a bias toward high-priority samples, which changes this distribution uncontrollably. It can be corrected by using importance-sampling (IS) weights. This importance-sample weight is annealed from one starting value β_0 to 1, which means that its affect is felt more strongly at the end of the stochastic process as the unbiased nature of the updates in RL is most important near convergence.[54]

In Ref. 34, Hessel *et al.* compare and combine improvements to a new state-of-the-art DQL algorithm, they called *Rainbow*, which we will also use in this paper.

D. Agent-environment interface

In this work, we transfer the theoretical framework of an MDP to concrete applications in Earth system dynamics by using the agent-environment interface. In this context, the concept of the agent is solely defined by its action set. The action set can be regarded as a collection of possible measures the international community could use to influence the system's trajectory. The agent uses a DRL algorithm outlined in Sec. II B. Concerning the detection of sustainable governance policies, we are mostly interested in the final outcomes the agent has learned rather than in letting the agent make real-world decisions later on. To assess the feasibility of finding sustainable policies, we also investigate the learning process. In this work, we intend to test our framework in the context of Earth system models. We focus on a particular type of Earth system models, which has been termed "World-Earth models."[2,42] In World-Earth modeling, one tries to capture the coevolving dynamics between biophysical dynamics of the Earth system on the one hand and on the other hand the social and economic dynamics of

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies,
Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

the World community. Since optimizing welfare may lead to policies that are neither sustainable nor safe,[55] we are interested in governance policies whose resulting trajectories stay within certain *"sustainability boundaries"* of the state space. These include both planetary and socioeconomic boundaries. We set up the environments based on Kittel *et al.*[21] and Nitzbon, Heitzig, and Parlitz,[56] both using deterministic nonlinear World-Earth models including planetary boundaries and social foundations. The dynamics are described by a set of coupled autonomous differential equations that define a continuous state space. In our setting, time is discretized in integration steps dt. At each nth step, the environment's dynamics are numerically solved (i.e., integrated) for a single time step $\mathrm{d}t_n = t_n - t_{n-1}$. Therefore, the environments fulfill the Markov property of the MDP.

A scheme of how this framework is implemented is shown in Fig. 1. In the following paragraphs, we provide more details on how we map the required parts for an MDP (i.e., concrete states in the environment, actions, and reward function) to World-Earth models. We conclude this section with some technical notes about implementation and hyperparameter search.

a. Environment 1: The AYS model. This model is a low-complexity model in three dimensions studied in Ref. 57 and described in more detail by Kittel *et al.*[21] It includes parts of climate change, welfare growth, and energy transformation. As compared to classical Earth system models, the AYS model is adapted. For simplicity, carbon dynamics A is not modeled in an explicit carbon cycle but assumed to follow an exponential relaxation toward equilibrium. The relation of the wealth of a society is modeled through the economic output Y, where the economy is assumed to have a constant basic growth rate. A renewable energy source with learning by doing dynamics is implemented via a renewable energy knowledge stock S. The state the agent observes at time t is, therefore, given by the tuple $s_t = (A, Y, S)_t$, consisting of three numerical values. As sustainability boundaries, we use a planetary boundary $A > \bar{A} = 345\,\mathrm{GtC}$ and a social foundation boundary $Y > \bar{Y} = 4 \cdot 10^{13}\ \$/yr$. For details, we refer to the Appendix or Ref. 21.

b. Environment 2: The copan:GLOBAL model. This model, studied by Nitzbon, Heitzig, and Parlitz,[56] is a conceptual model that describes the coevolution of natural and economic subsystems of the Earth. The model is meant for qualitative understanding of the complex interrelations rather than for quantitative predictions. Climate is represented as a global carbon cycle involving stocks of terrestrial carbon L, atmospheric carbon A, and geological carbon G, which influence the global mean temperature T. On the other hand, socioeconomic concepts, expressed through population P with capital K, are used to describe the flows of biomass and fossil fuels between society and nature. In Ref. 56, the authors consider a scenario where renewable energy does not exist. We extend the model for this study by including renewable energy use via a learning-by-doing dynamics for the renewable energy knowledge stock S, in the same manner as it was done in Ref. 42 for a regionalized version of Ref. 56. The state s_t is thus determined by the tuple $s_t = (L, A, G, T, P, K, S)_t$. Similarly, we use again $A > \bar{A} = 345\,\mathrm{GtC}$ and a social foundation boundary for consumption of $W > W_{SF} = 7850\,\$/\mathrm{yr}$ per capita as sustainability boundaries. For details of the system dynamics, the reader is referred to the Appendix or Refs. 42 and 56.

c. Action set. The action set $\mathcal{A}$ represents certain governance management options. It consists of no extra management (called default), a carbon tax, subsidies of renewable energies, for the c:GLOBAL environment of a nature protection policy, and all possible combinations of these management options. Depending on the specific environment, each action alters the dynamics of the state variables. For details, we refer to the Appendix.

d. Reward function. Reward functions express the agent's preferences over state-action pairs and, therefore, control the learning process. The reward functions are not a system feature but a parameter of the learning algorithm. We are free in our choice of the reward function and are guided in this choice by how well the chosen reward function helps the learner to achieve the actual goal. Since our ultimate objective is not to maximize some objectively given reward function but to stay within the boundaries, we chose the reward functions accordingly. Reward functions can be both continuously and discontinuously changing. To prove our framework working for both types, we used the following simple reward functions:

- *Survival reward*: provide a reward of 1 if the state s_t is within the boundaries, else 0.
- *Boundary distance reward*: calculate the distance of the state s_t to the boundaries in units of distances from the current state of the Earth to the boundaries. This distance is provided as a reward.

Depending on the chosen reward function, the trajectories found by the agent differ. In the case of survival reward, the agent is only interested in staying within the boundaries, whereas, in the latter case of the boundary distance reward, the agent tries to detect trajectories resulting in a large distance to the boundaries.

e. Implementation. After the experience, replay memory is filled with experiences from an agent that acts randomly in the environment. The learning process runs as follows. The agent is trained for a fixed number of episodes. A start position within the boundaries is randomly drawn from a uniform distribution of states around the current state. The number of iteration steps during one single learning episode is limited to a maximum of T. The end of one learning episode is determined either when T is reached or ended prematurely at time t either when a boundary is crossed or when approximate convergence to a fixed point is detected. In the latter case, the remaining future rewards are estimated with a discounted reward sum for the remaining time $T - t$ of the reward r_t. In any case, after the end of a learning episode, the environment is reset to time $t = t_0$ and a new start point s_{t_0} within the boundaries of the environment is randomly drawn.

f. Hyperparameter tuning. For each environment, we trained a different network. To get an optimal hyperparameter set for every environment, we systematically investigated the effect of various parameters on the learning success, such as the discount factor, the

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

training data mini-batch size, or the exploration-exploitation trade-off based on own exploration and on standard parameters of the DRL community as presented in Ref. 34. For a detailed explanation, we refer to the Appendix. The exploration rate ϵ starts with 1 and decays over time. We achieved the best performance for a replay buffer (i.e., the memory size) of 10^5 which is less than the default value in many DRL algorithms (e.g.,[25,34,52]) but in accordance with the work of Ref. 58 stating that the size of the replay buffer is crucially environment-dependent and needs a careful tuning. A full list of all hyperparameters can be found in Table I in the Appendix.

The neural network is based on the following architecture. The input layer of the size equaling the dimension of the state space is followed by two fully connected hidden layers, each one consisting of 256 units. The output layer is a fully connected linear layer that provides an output value for each possible action in the action set, representing the estimated Q-value of that action for the state given by the inputs. For minimizing the loss function, instead of simple stochastic gradient descent (SGD), the Adam optimizer[59] is used due to its better performance than SGD in DRL applications, as reported in Ref. 34.

III. APPLICATION TO WORLD-EARTH MODELS

Based on our proposal outlined above, we implemented an agent that learns by using a DRL (see Sec. II B) to manage the environments described in Sec. II D. The agent is trained for a maximum number of 10^4 episodes, where the learning success is evaluated every 50 episodes. Single simulation experiments can be carried out on standard notebook computers in a reasonable computing time (one to two hours on a single machine). Using a tuned hyperparameter set (see Table I in the Appendix for details), we can formulate three key findings of this work that is outlined below. First, we find that learning in terms of increasing rewards in the environments is indeed possible. Second, we investigate the specific pathways found by the learner and observe that the agent acts with great farsightedness. Moreover, we see a general strategy behind the detected trajectories that the learner has developed. Third, we explore that the agent also achieves good performance in scenarios in which the state space is only partially observable to the agent.

A. Training and stability

In order to verify the overall applicability of our algorithm, we first analyze the learning behavior in general. Unlike in supervised learning, where one can evaluate the performance of an algorithm by evaluating it on a set of test data, it is not obvious how to evaluate accurately the training progress an agent makes in RL problems. Here, we stick to the method used by Mnih *et al.*[25] visualizing the training properly. We plot the total reward the agent collects during one run over the number of learning episodes. Each value is computed as a running average over 50 episodes. Each curve is the average of 100 independent simulations.

As a result, we see that, after a certain number of episodes, the average reward per episode significantly increases in our environments (see Fig. 2). Obviously, the agent finds trajectories that reveal a high reward. In other words, it learns to manage the environment. We conclude that management can indeed be learned by the agent.

Furthermore, we observe that the learning of the agent is stabilized by using the extensions to DQN-Learning as outlined in Sec. II C. The plot suggests that the usage of dueling network architectures combined with double-Q-learning (DDQN + Duel) and prioritized experience replay with importance sampling (PER IS) significantly increases the performance of our DQN-Agent. The positive effect of PER IS can be explained by the observation that, in both environments, we find states in the resulting trajectories, where the dynamics significantly changes (as it will be outlined below). Experiences containing these states will be privileged in the learning process by PER IS. This is in good agreement with the results in Ref. 52. Therefore, all results outlined below are achieved by using our best performing agent (DDQN + Duel + PER IS), if not stated otherwise. Moreover, this is in qualitatively good accordance with other comparisons of different learning architectures, as, e.g., presented in Ref. 34, and the learning curves show a similar shape as seen in other DRL applications.[24,25,34]

B. Management pathways in World-Earth system models

In the following, we discuss the pathways in the two environments described in Sec. II D that were found by using the outlined framework of DRL. In Sec. III B 1, we explore the AYS model and, in Sec. III B 2, the copan:GLOBAL model. Specifically, in both

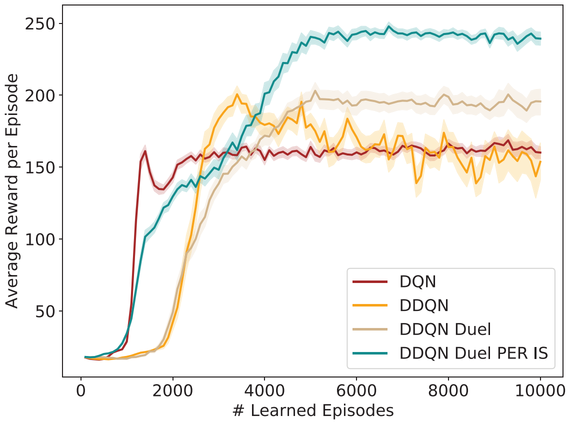


FIG. 2. Development of total average reward per episode. The average reward is a running average over the last 50 episodes of the sum of all rewards gained during one training episode. The curves are an average of 100 independent simulations of the AYS model. The light bands show 95% confidence intervals for the expected values estimated by these averages. Different deep-Q-network architectures are analyzed: DQN = deep-Q-networks, DDQN = Double DQN, DDQN Duel = Dueling network architecture with DDQN, DDQN Duel PER IS = DDQN Duel using prioritized experience replay together with importance sampling. The simulations were performed with a ϵ-greedy policy with ϵ decaying exponentially from 1 to 0.01 at a decay rate of $\lambda = 0.001$.

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

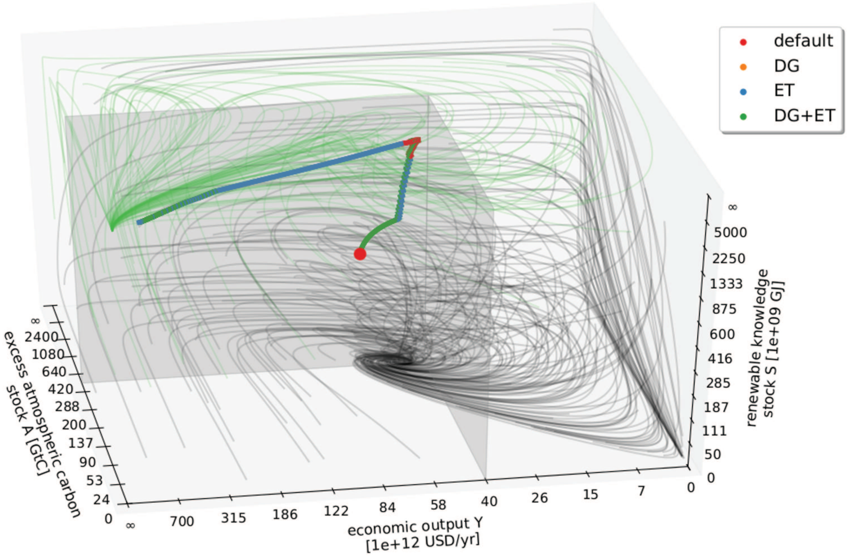


FIG. 3. Dynamics of a stylized World-Earth system according to Kittel *et al.*[21] The default flow of the AYS model is sampled with thin trajectories with randomly distributed initial conditions. We used nonlinearly scaled axes to account for the full space $\mathbb{R}^3$. The red dot in the center: Estimated current state of the world. Green lines: attraction basin of sustainable fix point which can be understood as the safe and just operating space. Black lines: attraction basin of the carbon-based economy. Gray surfaces: sustainability boundaries. In color, the example trajectory from the current state into a green future. The possible management options of the action set are: DG: degrowth, and energy transformation, i.e., carbon tax + subsidies on renewables.

environments, we are interested in whether the learner finds trajectories toward regions, which we can associate with a safe and just operating space without violating sustainability boundaries. First, we present some successful examples. As a next step, we look at the specific trajectories in more detail, hoping to understand the general strategy the agent found to reach its aim (i.e., maximize the total reward).

1. Pathways in the AYS model

In the AYS model, the agent can choose between the following actions: "energy transformation" (taxing carbon emissions and/or subsidizing renewables) or "degrowth management" (reducing the basic economic growth rate) or neither or both of them. As a first analysis step, we look at the pathways the agent takes after it was trained for a sufficiently long time (i.e., the convergence of the learning is reached, see Fig. 2). We find that, even though the dynamics of the environment is unknown to the agent in advance, it is able to find trajectories within sustainability boundaries (see Fig. 3) that were deemed impossible in another study based on a viability theory algorithm that used state space discretization.[21]

Due to the setup of our framework, each of the two management options can only be switched on and off. In Fig. 3, in the region near to the boundaries, the energy transformation (ET) option (representing an energy tax or subsidy) is switched on and off in short alternations, achieving essentially the effect that a continuous application of a smaller tax/subsidy would have. Hence, offering different tax levels as individual options might improve the learning success further.

To get a deeper understanding of the found solutions, we take a closer look at the different trajectories that were detected by using the DRL framework. Depending on the chosen reward function, the paths found by the agent differ. If the boundary distance reward is chosen, after sufficiently long learning, the agents always find a path toward the "green fixpoint" at $(A, Y, S) = (0, \infty, \infty)$, where the distance to the boundaries is maximized. For the survival reward, the agent is only interested in staying within the boundaries. Therefore, it finds pathways leading to the green fixpoint as well as pathways toward a region close to the boundaries with $S = 0$ where it then manages to stay. Although many viable paths are found by the learner, we find that the learning strategies that were found can be generalized. We analyzed the management options the agent uses most on different parts of the trajectories. They are depicted in Fig. 4. These different regions of predominant management options are now used for the following discussion. The different regions colored in Fig. 4 may be analyzed with respect to a mathematical theory of the qualitative topology of the state space of a dynamical system with management options and desirable states, called *topology of sustainable management* (TSM).[60] Interestingly, these regions can be seen to correspond roughly to some concepts from the TSM framework, in particular, the concept of "shelter" and "backwaters." The approximate locations of these regions are depicted by dashed lines in Fig. 4.

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies,
Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

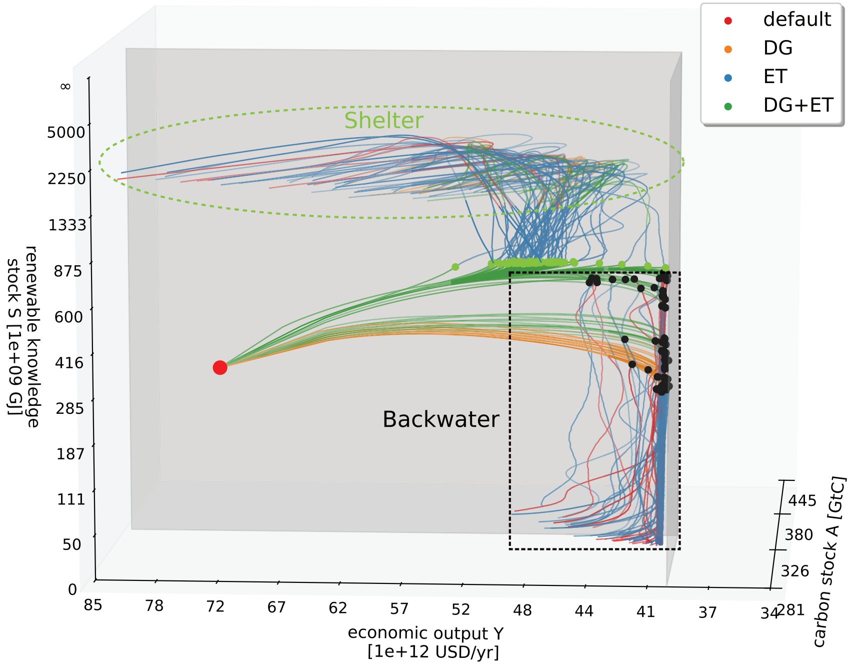


FIG. 4. Analysis of predominant management strategies in 200 independent simulations that find a trajectory inside the boundaries (gray surfaces). Half of them use the boundary distance reward and go toward the green fixpoint within the "shelter" region where management can be stopped (green dashed line). The others use the survival reward and go into a fossil-based future within the "backwaters" region from which no return to the shelter region is possible (black dashed line). Management options: DG: degrowth, reducing the basic growth rate of the economy; ET: energy transformation, taxing carbon and/or subsidizing renewables. Dots denote points of strongest gradients on each trajectory (green for going into the shelter and black for going into the backwaters). Here, the predominant learning strategy changes as well. The color of the trajectories shows the predominant management option used in each state. One can see that, close to the shelter, no specific management option is preferred since the choice becomes irrelevant.

We identify a general strategy the agent uses. Starting from the current state, we found that in order to stay within the boundaries forever, it is not sufficient to use only one single management option of energy transformation (ET) or degrowth (DG) in the beginning. Rather, both ET and DG have to be applied to ensure keeping the system within the sustainability boundaries in future times. To understand this behavior, one has to recall the effect of the two possible management options DG and ET (for details, we refer to the Appendix). Both boundaries $\bar{A}$ and $\bar{Y}$ are potentially dangerous for the learner. Using only option ET will lead to an increase of renewable knowledge but violate the $\bar{A}$ boundary. Applying option DG will on the one hand respect the $\bar{A}$ boundary but on the other hand hit the $\bar{Y}$ boundary. The strategy found by the learner is a mix of both options: First, it uses option ET + DG to reach a certain distance from the $\bar{A}$ boundary. However, the $\bar{Y}$ limit comes critically near. At a specific time point, the agent has to change its predominant management strategy to ET such that the renewable knowledge stock increases faster and the agents avoid transgressing the boundary value of $\bar{Y}$. At this specific point where DG + ET changes to ET, a sharp turn in the trajectory happens (see Fig. 4). If S is large enough at this point, the turn is "upwards" and after some time, a region is reached where every trajectory is now leading toward unlimited growth of economic output and renewable knowledge regardless of the chosen management option, so that management can be "stopped," leading to another sharp turn in the trajectories. In TSM, such a secure region is called a *shelter*. However, if S is too small at the turning point, the turn is "downward" toward $S = 0$, staying close to the social foundation boundary. In Ref. 21, it was shown that this leads to a region called the *backwaters,* from which the shelter could not be reached any longer, but one can still stay within the boundaries by managing over and over again.

Summarizing, the agent learns that the timing of the particular change of management is of crucial relevance. A general interpretation of the resulting pathways would be that ET, e.g., via taxing fossils, is highly important to ensure further development. However, to reach a secure state without violating the sustainability boundaries, a degrowth policy is needed for some time as well.

2. Pathways in the c:GLOBAL model

We verify that our framework works as well in higher-dimensional environments by applying it to the c:GLOBAL model. While classical approaches like viability theory are no longer well applicable because of the dimension, our DRL learner is also capable of detecting solutions toward a sustainable future in this model, see Fig. 5. Here, one learning episode has a maximum length of 500 yrs. Successful trajectories often converge already after around 100 yrs. However, to account for long-term effects, simulations were executed for times up to 500 yrs since we observed that seemingly converged trajectories sometimes transgressed boundaries at much later times, posing an additional challenge for the learner. The general strategy found by the learner turns out to be this. The NP option is used throughout and renewables are subsidized during

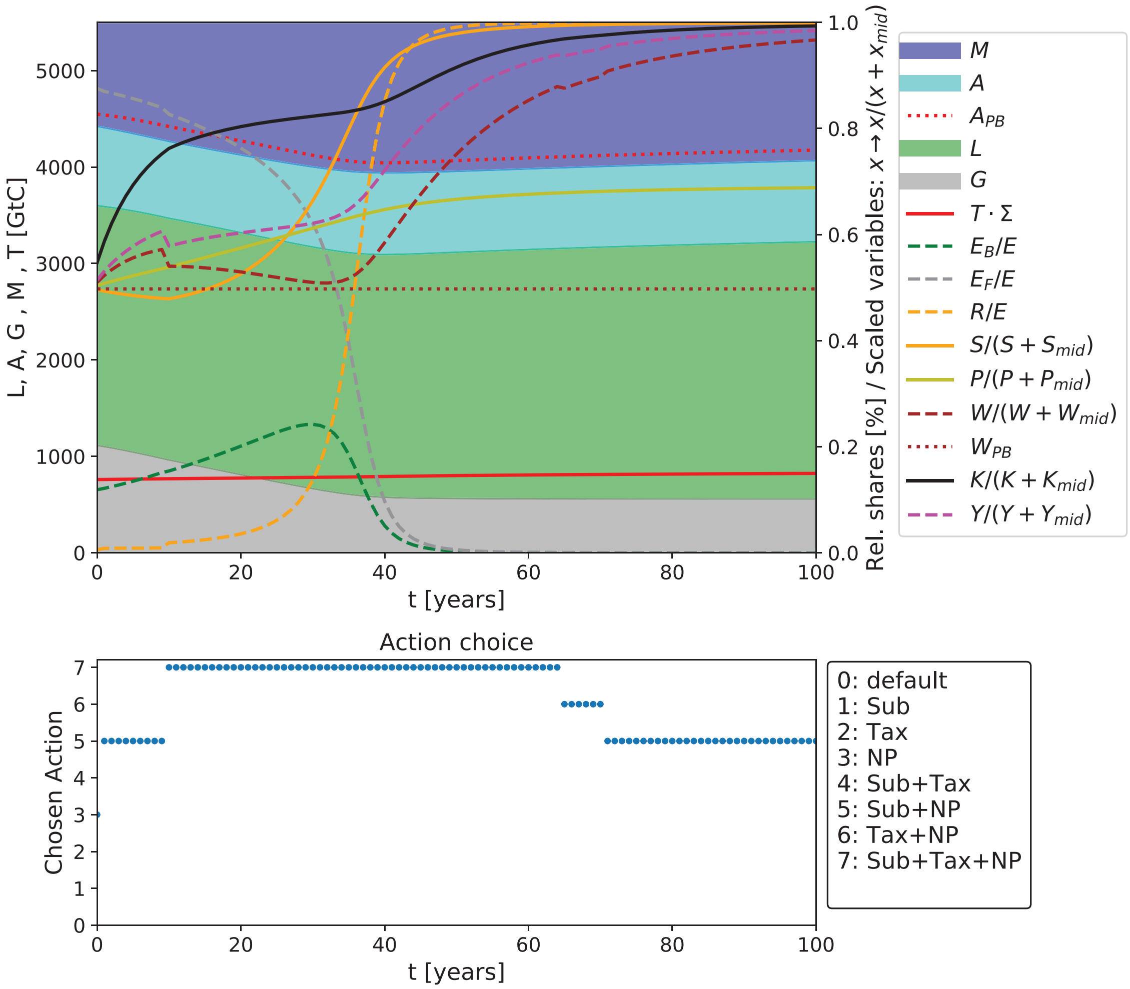


FIG. 5. Exemplary trajectories for successful management in the model based on Ref. 56. The upper graph shows the time evolution of the main variables, the lower the chosen actions at every time step. Dynamical variables are displayed as colored bands and solid lines, derived variables as dashed lines, and the planetary boundaries and social foundations in dotted lines. The total energy use is denoted as $E = E_R + E_B + R$. For visual reasons, we rescaled the S, P, W, K, Y with $S_{mid} = 5 \cdot 10^{11}$ bits, $P_{mid} = 6 \cdot 10^{9}$ H, $W_{mid} = W_{SF} = 7850\,\$/yr\,$H, $K_{mid} = 5 \cdot 10^{13}\,\$$, and $Y_{mid} = 6.3 \cdot 10^{13}\,\$/yr$. Since the system converges, only the first 100 years are shown. The available management options are Sub (subsidies on renewables), Tax (carbon tax on fossils), NP (nature protection for land use), and all combinations of these.

most of the time. The crucial point is the timing of the carbon tax, which cannot be used immediately without violating the social foundation boundary. It is switched on only later and switched off again once renewables have passed through most of their learning curve.

An interesting observation regarding the farsightedness of the agent is the following. After some learning episodes, the agent often uses trajectories that do not use any management during the years 20–60, which keeps the system within the boundaries for some time but leads to a violation of $\bar{A}$ later for some $t > 100$ yrs. One example trajectory can be found in Fig. 7 in the Appendix. Only after many more episodes, the agent learns to act with foresight and use management options early on that only make a recognizable difference much later and avoid crossing the boundaries. This is indeed a key feature for the success of DRL and shows the potential power of the method.

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

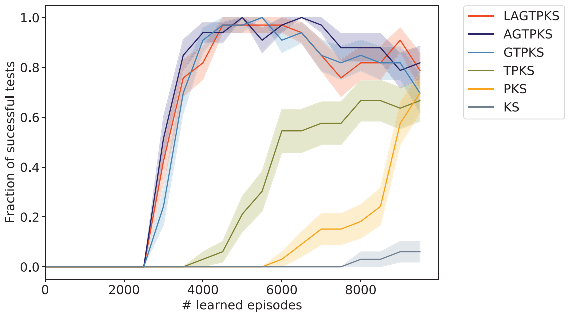


FIG. 6. Percentage of tests the agent passes successfully given different information about the state. An episode is considered successful if the agent, starting from the current state, manages to reach the shelter region where management can be turned off. For each set of dynamical variables, we simulated 100 independent learning processes and show 95% confidence bands for the reported percentages as estimates of the true success probabilities. The legend lists the variables observable by the agent: L = terrestrial carbon, A = atmospheric carbon, G = geological carbon, T = global mean temperature, P = population, K = capital, and S = renewable energy knowledge stock.

However, taking a look at the stability of the learning (see Fig. 6), we observe that the learning success in the copan:GLOBAL model also decreases again after a still larger number of episodes. As a possible explanation, we suggest that this is connected to the replay buffer. To avoid this phenomenon, the replay buffer needs to contain experiences, especially about the time steps where the dynamics of the system changes significantly.[58] After many successful runs, we still continue collecting observations in the memory buffer at every time step. Therefore, it mostly contains experiences for time points $t > 50$ yrs. However, especially the first time steps are crucial to avoid transgressing boundaries at later times as outlined above. These are, therefore, essential for the learning success. It seems that the agent tends to forget about experiences from early time steps and the learning success decreases. Further investigation considering the question of which experiences should be stored in the replay buffer could be the first step to overcome this issue.

C. Partial observability and noise

As a generalization of Markov decision processes, partially observable Markov decision processes (POMDPs) are of great research interest. Here, the agent is only able to observe only part of the actual system state.[61] We are interested in the performance of our DRL agent under such observational constraints since a real-world manager will only have access to vastly restricted information about the Earth system's current state. Moreover, we added noise to the observations of the agent. Our experiments show (see Fig. 6) that, even under partial observability of the state, the agent is still capable of detecting sustainable solutions. We observe that the learning curves for observing either the full state(L, A, G, T, P, K, S) or only the variable combinations (A, G, T, P, K, S) or (G, T, P, K, S) have very similar shape. So, it seems that there is little added value in observing the carbon stocks L and A when already observing the geological stock G whose decline is essential for the timing of the carbon tax (but which is also the hardest to observe in reality). However, even if we limit the agent's observation capabilities to the socioeconomic variables (P, K, S), the agent achieves a similar performance after a certain number of episodes, only considerably later. This can be explained by the dominant force humans exert on the Earth system.

To test the robustness of the DRL algorithm for a noisy state input, we added white observational noise on the input state s_t the agent receives from the environment. Not surprisingly, noise can disturb the agent's learning and lead to a massive decrease in performance if the environment gets more complex (see Fig. 8 in the Appendix for details). Neural networks are known to be vulnerable by perturbed input[62,63] and the harmful effect of noise has already been observed and discussed as well in DRL applications.[64–66] Still, for further experiments with more realistic scenarios, the influence of noise has to be investigated more systematically.

For the analysis of trajectories in the Earth system, we can deduce the following. Even if the full state will not be observable to the agent, it is just based on the distance boundary reward signal still able to sufficiently "understand" the system's dynamics in order to find appropriate management pathways. Furthermore, in our experiments, we see that noise will be a limiting factor for some DRL algorithms. In simulations with very noisy environments, some preprocessing of the input state might be necessary to use DRL successfully.

IV. CONCLUSION

The main contribution of this work is the development of a framework for using DRL in Earth system models, mathematically formalized in a Markov decision process. Throughout this paper, we have combined the technique of deep reinforcement learning with Earth system modeling in order to detect global sustainable management strategies. We have presented a prototype for which we hope extensions based on our work will become a helpful tool to discover and analyze management pathways and to get a deeper understanding of the impact of global governance policies.

As a proof of concept, we have applied it to two exemplary models from Earth system science, taken from the World-Earth modeling literature. They include components of Earth system modeling as well as constraints of planetary boundaries and social foundations. We have shown that our algorithm successfully identified trajectories toward a secure region for the Earth system which a competing approach using viability theory and a discretization of the state space were not able to find.[21] Even very simple reward functions were sufficient, and only partial observations of the system state were necessary for the learner to understand the complex, nonlinear system's dynamics. However, noisy observations have presented a challenge. We have found significant learning improvements by using the combination of DQN with dueling network architectures and prioritized experience replay and importance sampling.

With respect to management strategies that the learner found in the AYS and the c:GLOBAL model, we can support the intuition that there is not one single way for staying within the boundaries nor can the impact of global management be observed immediately. Rather, we conclude from our models that only an intelligent combination

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

and timing of global policies may lead to a sustainable future. We found that besides making renewables more attractive, also a temporary slowing down of economic growth might be necessary for staying within planetary boundaries.

Moreover, we have shown that our method is applicable as well in environments with only partially observable state spaces. Due to its connection to real-world problems,[61] for example, in 3D navigation,[67] partial observability of state spaces is widely discussed in the reinforcement learning community. Hence, in future work, the effects of reducing the dimensionality of the state space in our World-Earth system models need to be studied in more detail.

We used DRL to identify trajectories under certain constraints. Formally, this can be regarded as an optimization problem, which could be approached with other methods as well. For example, the IAM community typically uses commercial solvers for the optimization of long-term social welfare functions, which are influenced by nonlinear underlying dynamics. However, the choice of the welfare function is not directly intuitive and hard to justify straightforwardly.[16] As an example, Pindyck[16] puts forward the significant differences in the outcome of two established models in IAM. The results in Refs. 68 and 69 differ widely, mainly based on the different values of the discount rates for the choice of which no uniform theory exists. However, in our models, the constraints imposed by sustainability boundaries, as well as the choice of simple reward functions, could be argued to be easier to justify and to understand intuitively in some contexts.

We encourage the reader to apply our framework to his or her preferred models. Since we formulated our problem as an MDP, our approach is not restricted to deterministic environments but can be generalized to environments that include stochastic dynamics and agent-based components. One could think about replacing the global society used in the models above by agent-based models of regionally distributed interacting societies. Following the model developed by Wiedermann *et al.*,[70,71] which is a stochastic environment based on an adaptive network model, could be a first step in this direction. On the other hand, the biophysical dynamics could be incorporated in more detail as well by using more complex global vegetation models such as LPJ.[72]

Further, an interesting next step could be to use DQN agents to represent major real-world agents such as governments in a multiagent environment setting. Here, first experiments in simple grid worlds have already been performed to investigate sequential social dilemmas[73] and common-pool resource appropriation.[74] Connections to game theory in the climate context are conceivable as well.[47,75]

Another approach that might be promising is to include *model-based* RL in our framework. Regarding computation time, model-based RL tends to be much more efficient.[76] The key difference is that model-free methods act in the real environment in order to collect rewards and update the action value functions accordingly. In contrast, the agent in model-based methods uses RL to learn a model of the environment and then predicts the system dynamics in a second step. Once the model is learned, actions can be chosen by using optimal control theory. Specifically, as environments in World-Earth models are often based on a set of biophysical and socioeconomic differential equations, this approach might be promising. However, highly complex environments often cannot be learned perfectly, such that solutions of this method involve the risk of being suboptimal. A possible approach to overcome this issue is recently developed algorithms that aim to combine advantages of both methods in one algorithm.[77]

Another fruitful exchange could emerge between the field of Earth system analysis and the field of safe and beneficial AI.[78] For example, the important question of the latter field of how self-learning agents can safely explore an environment without pursuing catastrophic action directly translates to finding sustainable policies in Earth system analysis. Here as well, management strategies need to navigate uncertain environments without activating tipping elements in the Earth system with potentially catastrophic impacts on human societies.[79,80]

ACKNOWLEDGMENTS

This work was developed in the context of the COPAN collaboration at the Potsdam Institute for Climate Impact Research (PIK). The authors thank the COPAN group for helpful discussions and comments. Moreover, F.M.S. thanks the participants and organizers of the Eastern European Machine Learning (EEML) Summerschool for inspiring suggestions to this work. We are grateful for financial support by the European Research Council (via the ERC advanced grant project ERA), the Stordalen Foundation (via the Planetary Boundaries Research Network PB.net), the Earth League's EarthDoc program, and the Leibniz Association (project DOMINOES). The authors gratefully acknowledge the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research, and the Land Brandenburg for providing resources on the high-performance computer system at PIK.

APPENDIX A: THE AYS MODEL ENVIRONEMENT

In this environment, the observable state is composed of three real-valued components: the excess atmospheric carbon stock over preindustrial levels $A \geq 0$ (GtC), the gross world economic product $Y \geq 0(\$/yr)$, and the global knowledge stock for producing renewable energy $S \geq 0$ (GJ). The time evolution of these is given by three ordinary differential equations in which several additional, derived quantities occur that the agent cannot directly observe. These auxiliary variables are total world demand for primary energy U (GJ/yr), a relative price level of renewables G, resulting in a division of U into renewable energy production R and a flow of fossil energy F, and finally the global greenhouse gas emissions flow E (GtC/yr) resulting from F.

a. Dynamics in the AYS model. We assume that each unit of output Y requires a fixed amount $1/\epsilon$ of energy and the two energy sources are used in proportion to relative price (see Ref. 21 for a justification), so that

$$U = Y/\epsilon, \quad F = GU, \quad R = (1 - G)U, \quad E = F/\phi. \qquad \text{(A1)}$$

We assume the absolute price of fossils to remain constant and that of renewable energy to depend on the renewable knowledge stock in a power law relationship, so that the relative price of renewables vs

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

fossils has the form

$$G = \frac{1}{1 + (S/\sigma)^{\rho}}. \tag{A2}$$

Instead of assuming a carbon cycle as in the c:GLOBAL model (see below), we here simply assume that atmospheric carbon stock declines exponentially toward its equilibrium value where excess atmospheric carbon vanishes, so that

$$dA/dt = E - A/\tau_A.$$

Likewise, instead of assuming a classical economic growth model as in c:GLOBAL, we here simply assume that the gross world product grows at a fixed basic rate which is reduced in proportion to A (interpreted as a proxy for climate damages),

$$dY/dt = (\beta - \theta A)Y.$$

Finally, learning-by-doing makes the renewable knowledge stock grow with renewable energy production, and forgetting makes it decline exponentially,

$$dS/dt = R - S/\tau_S.$$

We use the following initial conditions and parameter estimates from Ref. 21: energy efficiency $\epsilon = 147\,\$\mathrm{GJ}^{-1}$, fossil combustion need $\phi = 4.7 \cdot 10^{10}\,\mathrm{GtCGJ}^{-1}$, break-even level of renewables $\sigma = 4 \cdot 10^{12}\,\mathrm{GJ}$, learning-by-doing exponent $\rho = 2$, characteristic time of natural carbon uptake $\tau_A = 50\,\mathrm{yr}$, basic economic growth rate $\beta = 0.03\,\mathrm{yr}^{-1}$, climate damage coefficient $\theta = 8.57 \cdot 10^{-5}\,\mathrm{yr}^{-1}$ GtC^{-1}, and characteristic time of forgetting $\tau_S = 50\,\mathrm{yr}^{-1}$.

b. Desirable region and management options. The AYS environment is an interesting minimum-complexity toy model for sustainability science because one can represent both the climate change planetary boundary and a wellbeing social foundation boundary in it by studying whether A may stay below some threshold $\bar{A} = 345\,\mathrm{GtC}$, and Y does not drop below some minimum value $\bar{Y} = 4 \cdot 10^{13}\,\$\,\mathrm{yr}^{-1}$ at the same time. In this paper, we assume the agent that represents the world community will try to avoid that the system converges to a fixed point with $S = 0$, $A > \bar{A}$, and $Y < \bar{Y}$, e.g., by making it instead go to $A = 0$ and $S, Y = \infty$ without violating the bounds. To do so, the agent has in this setup four options:

- **DG**: This option reduces the basic growth rate to $\bar{\beta} = \beta/2$, which helps to respect the boundary $\bar{A}$ but now risks to violate the social foundations boundary $\bar{Y}$.
- **ET**: This option supports an energy transition. It lowers the break-even point to $\bar{\sigma} = \sigma \cdot (1/2)^{\rho}$ which can be understood as subsidizing renewables and/or taxing fossils. This option does not change the location of the fixpoints but changes the dynamics significantly toward the green fixpoint.[21]
- **Default**: The agent can also use neither of these options.
- **DG + ET**: The combination of both of these options is possible as well.

APPENDIX B: THE c:GLOBAL MODEL ENVIRONEMENT

The model underlying this environment is of a similar type but more complex, having seven dynamic variables, of which the agent can observe different subsets in our experiments, as well as several additional unobserved auxiliary variables.

a. Dynamics in the c:GLOBAL model. Here, terrestrial carbon stock changes due to temperature-dependent photosynthesis (first term) and respiration (second term), and due to harvesting of biomass B,

$$dL/dt = (l_0 - l_T T)\sqrt{A/\Sigma}\, L - (a_0 + a_T T)L - B.$$

Absolute atmospheric carbon stock A changes due to photosynthesis, respiration, combustion of harvested biomass ($= -dL/dt$), and ocean-atmosphere diffusion,

$$dA/dt = -dL/dt + \delta(M - mA).$$

Geological carbon stock G declines because of extraction of fossil fuels F,

$$dG/dt = -F.$$

Global mean temperature converges to a value dependent on A due to the greenhouse effect and is hence measured for simplicity on a nonlinear scale in units of atmospheric carbon per land surface area, so that

$$dT/dt = g(A/\Sigma - T).$$

Population P has a fertility (first term) and mortality (second term) that depend on wellbeing W,

$$dP/dt = P\left(\frac{2WW_p}{W^2 + W_p^2}\, p - \frac{q}{W}\right).$$

Physical capital K grows since part of GWP Y is invested and decays exponentially,

$$dK/dt = iY - kK.$$

For renewable knowledge stock S, we assume the same dynamics as in the AYS model,

$$dS/dt = s_R R - s_S S.$$

Since total carbon is fixed at C^*, maritime carbon stock M is

$$M = C^* - L - A - G.$$

Usage of the three assumed perfectly substitutable energy forms of biomass B, fossil F, and renewable energy flow R is determined by a general price equilibrium model (see Refs. 42 and 56) that leads to these equations,

$$B = \frac{a_B}{e_B}\,\frac{L^2 (PK)^{2/5}}{(a_B L^2 + a_F G^2 + a_R S^2)^{4/5}}, \tag{B1}$$

$$F = \frac{a_F}{e_F}\,\frac{G^2 (PK)^{2/5}}{(a_B L^2 + a_F G^2 + a_R S^2)^{4/5}}, \tag{B2}$$

$$R = a_R \frac{S^2 (PK)^{2/5}}{(a_B L^2 + a_F G^2 + a_R S^2)^{4/5}}. \tag{B3}$$

Economic output is proportional to energy input,

$$Y = y_E(e_B B + e_F F + R).$$

Finally, wellbeing is determined by per capita consumption and ecosystem services, which are assumed proportional to terrestrial carbon density,

$$W = \frac{(1-i)Y}{P} + w_L \frac{L}{\Sigma}.$$

We use the following initial conditions and parameter estimates from Refs. 42 and 56, which are based on data from year 2000: initial values $L_0 = 2480$ GtC (GtC = gigatons carbon), $A_0 = 830$ GtC, $G_0 = 1125$ GtC, $T_0 = 5.05 \cdot 10^{-6}$ GtCm^{-2} (global mean surface air temperature is not measured in Kelvin but for simplicity in carbon-equivalent degrees, i.e., GtC), $P_0 = 6 \cdot 10^9$ H (H = humans), $K_0 = 5 \cdot 10^{13}$ \$, and $S_0 = 5 \cdot 10^{11}$ bits.

The parameters are as follows: total available carbon stock $C^* = 5500$ GtC; photosynthesis parameters $l_0 = 26.4$ km yr^{-1} GtC$^{-1/2}$ and $l_T = 1.1 \cdot 10^6$ km^3 yr^{-1} GtC$^{-3/2}$; total land mass $\Sigma = 1.5 \cdot 10^8$ m^2; respiration parameters $a_0 = 0.03$ yr^{-1} and $a_T = 3200$ km yr^{-1}GtC$^{-1/2}$; diffusion coefficient $\delta = 0.01$ yr^{-1}; solubility coefficient $m = 1.5$; strength of the greenhouse effect $g = 0.02$ yr^{-1}; peak fertility wellbeing level $W_p = 2000$ \$yr^{-1} H^{-1}; peak fertility $p = 0.04$ yr^{-1}; mortality coefficient $q = 20$ \$yr^{-2}; savings and capital depreciation rates $i = 0.25$ and $k = 0.1$ yr^{-1}; knowledge accumulation $s_R = 1.0$ bits GtC^{-1}; forgetting parameter $s_S = 1/50$ yr^{-1}; total carbon $C^* = 5500$ GtC; energy subsector productivities $a_B = 1.5 \cdot 10^4$GJ^{5}yr^{-5}GtC^{-2}\$$^{-2}$H^{-2}, $a_F = 2.7 \cdot 10^5$GJ^{5}yr^{-5}GtC^{-2}\$$^{-2}$H^{-2}, and $a_R = 9 \cdot 10^{-15}$GJ^{5}yr^{-5}bits^{-2}\$$^{-2}$H^{-2}; energy efficiencies $e_B = 4 \cdot 10^{10}$ GJGtC^{-1} and $e_F = 4 \cdot 10^{10}$ GJGtC^{-1}; final sector productivity $y_E = 120$ \$GJ^{-1}; and wellbeing-sensitivity on ecosystem services $w_L = 0$ km^2 GtC^{-1} yr$^-$ H^{-1}

b. Desirable region and management options. The c:GLOBAL model allows us to include both planetary boundaries and the social foundations. To be consistent with the AYS model (see above), we use the same value for the state variable $\bar{A} = 945$ GtC (note that here A describes the amount of total atmospheric carbon, in contrast to the AYS model where A describes the excess atmospheric carbon after the beginning of the industrial revolution). However, in contrast to the AYS Environment, the social foundations boundary is not included directly through a dynamical state variable but comes within a derived variable. Since we have in the c:GLOBAL model a direct measure for wellbeing W, we use it as a boundary for the social foundations. The value is chosen as in the AYS model derived from economic production in year 2000 and set to $\bar{W} = 7850$ \$H^{-1}. Due to the mass conversation law, in any case, the dynamical variables L, A, and G cannot exceed the total amount of carbon C^* at any time. However, we find that the state variables P, K, and S diverge for large times with our chosen initial conditions. Accordingly, we define the desirable region that should be reached within this framework as a region that is within the planetary boundaries of $\bar{A}$, with still growing wellbeing (i.e., $W \to \infty$) and a population of above $1 \cdot 10^{10}$ H since this is a realistic estimation for the growth of the world population. As we saw examples where pathways fulfilled these requirements, but hit the planetary boundaries at some later time steps, we demanded furthermore that the agent has to stay in this region for a time of $t > 400$ yr. To find a pathway in these regions, the agent has in this setup eight different management options:

- **Sub**: This option doubles the energy availability of renewable energies via the factor a_R in Eq. (A5), which can be understood as a subsidy on renewables. It helps to push the development of renewable energy knowledge S but has no direct influence on the carbon flow.
- **Tax**: This option describes a carbon tax. It lowers the carbon based energy availability via the factors a_G and a_F in Eqs. (A4) and (A3), respectively. The factors a_F and a_B are decreased by a factor of 50%, which is roughly the price of carbon proposed in the last IPCC landmark report.[10] This option increases the distance to the planetary boundary $\bar{A}$ but lowers significantly the distance to the social foundations, especially at the beginning of the simulation.
- **NP**: Nature protection policy. It reduces the amount of terrestrial bound carbon L that can be used for energy generation to a proportion of 70% of the initial terrestrial carbon L_0. This option helps to avoid that too much terrestrial carbon is set free to the atmosphere but has no influence on the flow of geological carbon into the atmosphere and risks furthermore the probability to violate the social foundations $\bar{W}$.
- **Default**: The agent can also use neither of these options.
- **Combinations of Sub + Tax + NP**: All four possible combinations of the three single options can be used as well.

APPENDIX C: UNSUCCESSFUL MANAGEMENT IN c:GLOBAL

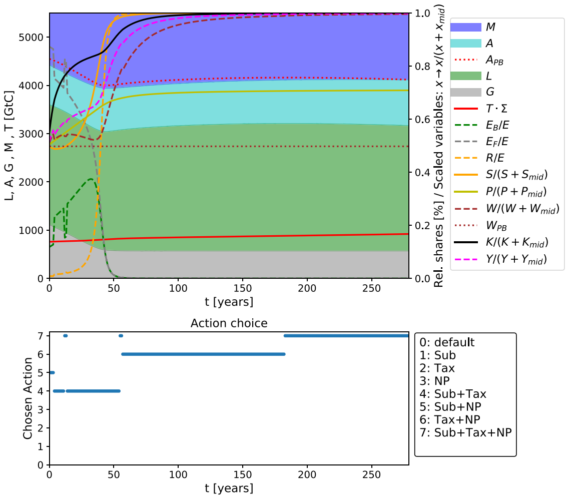


FIG. 7. Exemplary trajectories for unsuccessful management in the model based on Ref. 56. The upper graph shows the different trajectories, and the lower shows the chosen action at that time. Dynamical variables are displayed in solid lines, derived variables in dashed lines, and planetary boundaries and Social Foundations in dotted lines. The total energy use is denoted as $E = E_R + E_B + R$. For visual reasons, we rescaled the S, P, W, K, Y with $S_{mid} = 5 \cdot 10^{11}$ bits, $P_{mid} = 6 \cdot 10^9$ H, $W_{mid} = 7850$ \$/aH, $K_{mid} = 5 \cdot 10^{13}$ \$, and $Y = 6.2 \cdot 10^{13}$ \$/a. We plotted up to that time point when the $\bar{A}$ boundary is hit. The available management options were as follows: Sub = Subsidies on renewables, Tax = Carbon tax on fossils, and NP = Nature protection for landuse.

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

APPENDIX D: NOISY INPUT TO ENVIRONMENTS

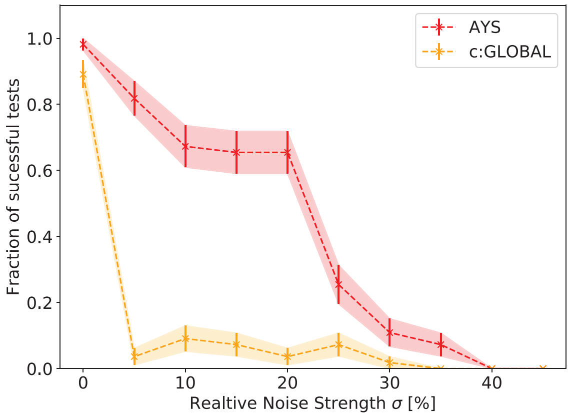


FIG. 8. Percentage of successful tests for environments with different levels of noise strength σ. White noise is set on the input states, and the strength of the noise could be up to σ relative to the input state s_t. We compare the learning success in the AYS as well as in the c:GLOBAL model. The agent is provided in all dimensions of the state space.

APPENDIX E: LIST OF HYPERPARAMETERS

In Table I, the list of the hyperparameters in AYS and c:GLOBAL environment is shown. The hyperparameter search was mainly done based on own exploration. Due to high computational costs, no systematic grid search was performed, but as one parameter was tested, the remaining were fixed at their previously explored optimal values. For the priority of transition α, the initial importance sampling weighting β_0, and the Adam optimizer learning rate, the recommended values in Refs. 54 and 34 were used.

TABLE I. Hyperparameters for the AYS and the c:GLOBAL environment.

Hyperparameter	Value AYS	Value c:GLOBAL	Description
Batch size	32	32	Number of training observations over which Q-value function update is computed
Replay memory size	$1 \cdot 10^5$	$1 \cdot 10^5$	Number of stored observations in replay memory
Initial exploration	1	1	Initial value in ϵ-greedy policy
Final exploration	0.01	0.001	Final value in ϵ-greedy policy
Decay rate exploration	0.001	0.001	Exponential decay of ϵ toward final value
Target network update frequency	100	200	The number of episodes after which the parameters of the target network are updated to the current network parameters
Adam learning rate	0.000 25	0.000 25	The initial learning rate in Adam optimizer
Discount factor γ	0.96	0.96	Discount factor used in Q-learning update
Priority of transition α	0.6	0.6	In prioritized experience replay: The exponent determines how much prioritization is used
Initial importance sampling weight β_0	0.4	0.4	In importance sampling β is annealed from β_0 to 1, which means that its affect is more relevant at the end of the simulation

REFERENCES

[1] H. J. Schellnhuber, "'Earth system' analysis and the second copernican revolution," Nature **402**, C19 (1999).

[2] J. F. Donges, R. Winkelmann, W. Lucht, S. E. Cornell, J. G. Dyke, J. Rockström, J. Heitzig, and H. J. Schellnhuber, "Closing the loop: Reconnecting human dynamics to earth system science," Anthropocene Rev. **4**, 151–157 (2017).

[3] J. Rockström, W. Steffen, K. Noone, Å. Persson, F. S. Chapin III, E. F. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber *et al.*, "A safe operating space for humanity," Nature **461**, 472 (2009).

[4] J. Rockström, W. L. Steffen, K. Noone, Å. Persson, F. S. Chapin III, E. Lambin, T. M. Lenton, M. Scheffer, C. Folke, H. J. Schellnhuber *et al.*, "Planetary boundaries: Exploring the safe operating space for humanity," Ecol. Soc. **14**, 32 (2009).

[5] UG Assembly, "Transforming our world: The 2030 agenda for sustainable development," Technical Report (Institution United Nations, 2015).

[6] UNFC on Climate Change, "Conference of the parties—Adoption of the paris agreement," Technical Report (Institution United Nations, 2015).

[7] J. M. Anderies, S. R. Carpenter, W. Steffen, and J. Rockström, "The topology of non-linear global carbon dynamics: From tipping points to planetary boundaries," Environ. Res. Lett. **8**, 044048 (2013).

[8] W. Steffen, K. Richardson, J. Rockström, S. E. Cornell, I. Fetzer, E. M. Bennett, R. Biggs, S. R. Carpenter, W. De Vries, C. A. De Wit *et al.*, "Planetary boundaries: Guiding human development on a changing planet," Science **347**, 1259855 (2015).

[9] K. Raworth, "A safe and just space for humanity: Can we live within the doughnut," Oxfam Policy Pract. Clim. Change Resil **8**, 1–16 (2012).

[10] J. Rogelj, D. Shindell, K. Jiang, S. Fifita, P. Forster, V. Ginzburg, C. Handa, H. Kheshgi, S. Kobayashi, E. Kriegler *et al.*, "Mitigation pathways compatible with 1.5 c in the context of sustainable development," IPCC Report (2018).

[11] W. Steffen, J. Rockström, K. Richardson, T. M. Lenton, C. Folke, D. Liverman, C. P. Summerhayes, A. D. Barnosky, S. E. Cornell, M. Crucifix, J. F. Donges, I. Fetzer, S. J. Lade, M. Scheffer, R. Winkelmann, and H. J. Schellnhuber, "Trajectories of the earth system in the anthropocene," Proc. Natl. Acad. Sci. U.S.A. **115**, 8252–8259 (2018), see https://www.pnas.org/content/115/33/8252.full.pdf.

[12] F. Müller-Hansen, M. Schlüter, M. Mäs, J. F. Donges, J. J. Kolb, K. Thonicke, and J. Heitzig, "Towards representing human behavior and decision making in earth system models—An overview of techniques and approaches," Earth Syst. Dyn. **8**, 977–1007 (2017).

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

[13] D. L. Kelly, C. D. Kolstad, *et al.*, "Integrated assessment models for climate change control," in *International Yearbook of Environmental and Resource Economics 1999/2000: A Survey of Current Issues* (Edward Elgar, Cheltenham, 1999), pp. 171–197.

[14] C. Pahl-Wostl, C. Schlumpf, M. Büssenschütt, A. Schönborn, and J. Burse, "Models at the interface between science and society: Impacts and options," Integr. Assess. **1**, 267–280 (2000).

[15] M. R. Bussieck and A. Meeraus, "General algebraic modeling system (GAMS)," in *Modeling Languages in Mathematical Optimization* (Springer, 2004), pp. 137–157.

[16] R. S. Pindyck, "The use and misuse of models for climate policy," Rev. Environ. Econ. Policy **11**, 100–114 (2017).

[17] M. I. Kamien and N. L. Schwartz, *Dynamic Optimization: The Calculus of Variations and Optimal Control in Economics and Management* (Courier Corporation, 2012).

[18] W. Liang, "Climate modification directed by control theory," e-print arXiv:0805.0541 (2008).

[19] N. Botta, P. Jansson, and C. Ionescu, "The impact of uncertainty on optimal emission policies," Earth Sys. Dyn. **9**, 525–542 (2018).

[20] G. Deffuant and N. Gilbert, *Viability and Resilience of Complex Systems: Concepts, Methods and Case Studies from Ecology and Society* (Springer Science & Business Media, 2011).

[21] T. Kittel, R. Koch, J. Heitzig, G. Deffuant, J.-D. Mathias, and J. Kurths, "Operationalization of topology of sustainable management to estimate qualitatively different regions in state space," e-print arXiv:1706.04542 (2017).

[22] R. S. Sutton and A. G. Barto, *Introduction to Reinforcement Learning* (MIT Press, Cambridge, 1998). Vol. 135.

[23] F. B. von der Osten, "Intelligent decision-making in coupled socio-ecological systems," Ph.D. thesis (University of Melbourne, 2017).

[24] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," Nature **518**, 529 (2015).

[25] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," e-print arXiv:1312.5602 (2013).

[26] G. Tesauro, "Temporal difference learning and TD-Gammon," Commun. ACM. **38**, 58–68 (1995).

[27] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," e-print arXiv:1708.05866 (2017).

[28] Y. Li, "Deep reinforcement learning," e-print arXiv:1810.06339 (2018).

[29] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. thesis (King's College, Cambridge, 1989).

[30] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**, 436 (2015).

[31] L.-J. Lin, "Reinforcement learning for robots using neural networks," Technical Report (DTIC Document, 1993).

[32] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," Nature **529**, 484 (2016).

[33] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," Science **362**, 1140–1144 (2018), see https://science.sciencemag.org/content/362/6419/1140.full.pdf.

[34] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," in *Thirty-Second AAAI Conference on Artificial Intelligence* (AAAI Press, 2018).

[35] H. Mao, M. Alizadeh, I. Menache, and S. Kandula, "Resource management with deep reinforcement learning," in *Proceedings of the 15th ACM Workshop on Hot Topics in Networks* (ACM, 2016), pp. 50–56.

[36] Z. Zhou, X. Li, and R. N. Zare, "Optimizing chemical reactions with deep reinforcement learning," ACS. Cent. Sci. **3**, 1337–1344 (2017).

[37] T. P. Lillicrap, J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning (2015)," e-print arXiv:1509.02971 (2016).

[38] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," J. Mach. Learn. Res. **17**, 1334–1373 (2016).

[39] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2017), pp. 3357–3364.

[40] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2017), pp. 3389–3396.

[41] T. Haarnoja, A. Zhou, S. Ha, J. Tan, G. Tucker, and S. Levine, "Learning to walk via deep reinforcement learning," e-print arXiv:1812.11103 (2018).

[42] J. F. Donges, J. Heitzig, W. Barfuss, J. A. Kassel, T. Kittel, J. J. Kolb, T. Kolster, F. Müller-Hansen, I. M. Otto, M. Wiedermann *et al.*, "Earth system modelling with complex dynamic human societies: The copan:Core World-Earth modeling framework," Earth Syst. Dyn. Discuss. **2018**, 1–27.

[43] W. B. Arthur, "Designing economic agents that act like human agents: A behavioral approach to bounded rationality," Am. Econ. Rev. **81**, 353–359 (1991).

[44] E. Lindkvist and J. Norberg, "Modeling experiential learning: The challenges posed by threshold dynamics for sustainable renewable resource management," Ecol. Econ. **104**, 107–118 (2014).

[45] E. Lindkvist, Ö. Ekeberg, and J. Norberg, "Strategies for sustainable management of renewable resources during environmental change," Proc. R. Soc. B **284**, 20162762 (2017).

[46] D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown *et al.*, "Tackling climate change with machine learning," e-print arXiv:1906.05433 (2019).

[47] W. Barfuss, J. F. Donges, and J. Kurths, "Deterministic limit of temporal difference reinforcement learning for stochastic games," Phys. Rev. E **99**, 043305 (2019).

[48] M. Wiering and M. van Otterlo, "Reinforcement learning: State-of-the-Art,“ in *Adaptation, Learning, and Optimization* (Springer-Verlag, Berlin, 2012), Vol. 12, pp. 3–42.

[49] R. Bellman, "A Markovian decision process," J. Math. Mech. **6**, 679–684 (1957).

[50] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Thirtieth AAAI Conference on Artificial Intelligence* (AAAI Press, 2016).

[51] H. V. Hasselt, "Double Q-learning," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2010), pp. 2613–2621.

[52] Z. Wang, T. Schaul, M. Hessel, H. Van Hasselt, M. Lanctot, and N. De Freitas, "Dueling network architectures for deep reinforcement learning," e-print arXiv:1511.06581 (2015).

[53] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (JMLR.org, 2017), pp. 449–458.

[54] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," e-print arXiv:1511.05952 (2015).

[55] W. Barfuss, J. F. Donges, S. J. Lade, and J. Kurths, "When optimization for governing human-environment tipping elements is neither sustainable nor safe," Nat. Commun. **9**, 2354 (2018).

[56] J. Nitzbon, J. Heitzig, and U. Parlitz, "Sustainability, collapse and oscillations in a simple world-earth model," Environ. Res. Lett. **12**, 074020 (2017).

[57] J. Heitzig, W. Barfuss, and J. Donges, "A thought experiment on sustainable management of the earth system," Sustainability **10**, 1947 (2018).

[58] S. Zhang and R. S. Sutton, "A deeper look at experience replay," e-print arXiv:1712.01275 (2017).

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," e-print arXiv:1412.6980 (2014).

[60] J. Heitzig, T. Kittel, J. F. Donges, and N. Molkenthin, "Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the earth system," Earth Syst. Dyn. **7**, 21–50 (2016).

[61] M. T. Spaan, "Partially observable Markov decision processes," in *Reinforcement Learning* (Springer, 2012), pp. 387–414.

[62] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," e-print arXiv:1312.6199 (2013).

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

[63] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (IEEE, 2016), pp. 372–387.

[64] V. Behzadan and A. Munir, "Vulnerability of deep reinforcement learning to policy induction attacks," in *International Conference on Machine Learning and Data Mining in Pattern Recognition* (Springer, 2017), pp. 262–275.

[65] V. Behzadan and A. Munir, "Whatever does not kill deep reinforcement learning, makes it stronger," e-print arXiv:1712.09344 (2017).

[66] S. Huang, N. Papernot, I. Goodfellow, Y. Duan, and P. Abbeel, "Adversarial attacks on neural network policies," e-print arXiv:1702.02284 (2017).

[67] P. Mirowski, R. Pascanu, F. Viola, H. Soyer, A. J. Ballard, A. Banino, M. Denil, R. Goroshin, L. Sifre, K. Kavukcuoglu *et al.*, "Learning to navigate in complex environments," e-print arXiv:1611.03673 (2016).

[68] W. D. Nordhaus, "Estimates of the social cost of carbon: Background and results from the rice-2011 model," Technical Report (Institution National Bureau of Economic Research, 2011).

[69] N. Stern and N. H. Stern, *The Economics of Climate Change: The Stern Review* (Cambridge University Press, 2007).

[70] M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, "Macroscopic description of complex adaptive networks coevolving with dynamic node states," Phys. Rev. E **91**, 052801 (2015).

[71] W. Barfuss, J. F. Donges, M. Wiedermann, and W. Lucht, "Sustainable use of renewable resources in a stylized social–ecological network model under heterogeneous resource distribution," Earth Syst. Dyn. **8**, 255–264 (2017).

[72] S. Sitch, B. Smith, I. C. Prentice, A. Arneth, A. Bondeau, W. Cramer, J. O. Kaplan, S. Levis, W. Lucht, M. T. Sykes *et al.*, "Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model," Glob. Chang. Biol. **9**, 161–185 (2003).

[73] J. Z. Leibo, V. Zambaldi, M. Lanctot, J. Marecki, and T. Graepel, "Multiagent reinforcement learning in sequential social dilemmas," in *Proceedings of the 16th Conference on Autonomous Agents and Multiagent Systems* (International Foundation for Autonomous Agents and Multiagent Systems, 2017), pp. 464–473.

[74] J. Perolat, J. Z. Leibo, V. Zambaldi, C. Beattie, K. Tuyls, and T. Graepel, "A multiagent reinforcement learning model of common-pool resource appropriation," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2017), pp. 3643–3652.

[75] J. Heitzig, K. Lessmann, and Y. Zou, "Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games," Proc. Natl. Acad. Sci. U.S.A. **108**, 15739–15744 (2011).

[76] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2018), pp. 7559–7566.

[77] V. Pong, S. Gu, M. Dalal, and S. Levine, "Temporal difference models: Model-free deep RL for model-based control," e-print arXiv:1802.09081 (2018).

[78] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," e-print arXiv:1606.06565 (2016).

[79] T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber, "Tipping elements in the earth's climate system," Proc. Natl. Acad. Sci. U.S.A. **105**, 1786–1793 (2008), see https://www.pnas.org/content/105/6/1786.full.pdf.

[80] H. J. Schellnhuber, "Tipping elements in the earth system," Proc. Natl. Acad. Sci. U.S.A. **106**, 20561–20563 (2009), see https://www.pnas.org/content/106/49/20561.full.pdf.

Reproduced from F. M. Strnad et al., Deep reinforcement learning in World-Earth system models to discover sustainable management strategies, Chaos, vol. 29, no. 12, p. 123122, 2019, doi: 10.1063/1.5124673, with the permission of AIP Publishing.

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-04738-z OPEN

When optimization for governing human-environment tipping elements is neither sustainable nor safe

Wolfram Barfuss[1,2], Jonathan F. Donges[1,3], Steven J. Lade[3,4] & Jürgen Kurths[1,2,5]

Optimizing economic welfare in environmental governance has been criticized for delivering short-term gains at the expense of long-term environmental degradation. Different from economic optimization, the concepts of sustainability and the more recent safe operating space have been used to derive policies in environmental governance. However, a formal comparison between these three policy paradigms is still missing, leaving policy makers uncertain which paradigm to apply. Here, we develop a better understanding of their interrelationships, using a stylized model of human-environment tipping elements. We find that no paradigm guarantees fulfilling requirements imposed by another paradigm and derive simple heuristics for the conditions under which these trade-offs occur. We show that the absence of such a master paradigm is of special relevance for governing real-world tipping systems such as climate, fisheries, and farming, which may reside in a parameter regime where economic optimization is neither sustainable nor safe.

[1] Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany. [2] Department of Physics, Humboldt University, 12489 Berlin, Germany. [3] Stockholm Resilience Centre, Stockholm University, 11419 Stockholm, Sweden. [4] Fenner School of Environment and Society, The Australian National University, Canberra, ACT 2601, Australia. [5] Saratov State University, Saratov 410012, Russia. Correspondence and requests for materials should be addressed to W.B. (email: barfuss@pik-potsdam.de)

The Sustainable Development Goals[1] and the Paris climate agreement set the target of prosperous development for people and our planet. Yet, it remains challenging to translate these aims into concrete policy implementations, accounting for non-linearities, such as tipping elements[2,3], regime shifts[4,5], and multi-stabilities[6], as well as multiple kinds of uncertainties[7–9], and extreme events[10].

To support the decision making processes in these contexts, we ask the question how the three prominent decision making paradigms of economic welfare optimization, sustainability and safe operating space compare with each other. Specifically, we investigate the parameter regimes for synergies and trade-offs when applying these paradigms to the management of tipping elements[11] and how these findings relate to the three real-world systems of climate, fisheries and farming.

Optimization approaches have emerged as the primary guiding principle to derive a policy strategy for environmental governance[12,13]. Most often, the present value of macroeconomic social welfare, i.e., the sum of discounted future benefits minus costs, is the target to be optimized. Such optimization approaches have been criticized regarding the discount rates used, delivering short term gains at the expense of long-term environmental degradation[14,15]. Further criticism targets the lack of a systems perspective required to understand the structural landscape of model dynamics, as well as the assumptions made due to imperfect information[6,9,10]. This critique is partly dealt with in optimization variants, such as robust[7,16] or viable[17–19] control, which are dealing with multiple types of uncertainty[20]. Naturally, other or multiple objectives[21] and criteria[22,23] with possible constraints[24] can be optimized as well. In this work, we use the term solely in the narrow economic sense of maximizing the present value as defined in Eq. 1 below.

In recognition of increasing environmental and social threats[25] the policy paradigm of sustainability has emerged in the scientific and political discourse[26,27]. The economics of sustainability has brought up many definitions of sustainability alone[28–31]. In these analyses sustainability is usually imposed as a constraint within an economic welfare optimization paradigm. Trade-offs to economic welfare optimization are well known[28,32]. However, these classic social welfare optimization approaches are challenged through the increasing recognition of non-linearities, such as tipping points, regime shifts, uncertainties and the risk of catastrophic outcomes[6,9]. Taking up these challenges, e.g., non-convexities[33] and climate tipping elements[34,35] have been studied within an economic framework. Here, we derive our formal definition of sustainability from the Brundtland report[26]. Its design is deliberately simple and targeted to the mathematical framework we use (see below). We do not intend our definition to be applicable to a general model of a welfare economy[12,27].

Recent advances in sustainability science have brought forth tolerable windows[36] or safe operating spaces[37,38] as a policy paradigm to derive concrete actions from[39]. These concepts originate from resilience thinking[40] and a precautionary principle[41] to deal with potential dangerous tipping elements in the environmental governance system. Trade-offs but also synergies with optimization thinking have been therefore discussed[42]. Also formal analyses studying relations between resilience as a system property and sustainability were conducted[43,44].

However, the reciprocal relationships between these three paradigms of economic optimization, sustainability and safe operating space is still insufficiently explored. Such an understanding is important in order to judge, for example, when economic optimization is, or is not, an appropriate policy goal. Also, guidance is required when a sustainability paradigm may conflict with a safe operating space paradigm and vice versa.

Here, we report progress towards a better understanding of the mutual relationships between these three paradigms of economic optimization, sustainability and safe operating space by applying them to a stylized model of a human-environment tipping element. We do so because of the increasing importance of tipping points and regime shifts in environmental governance. Our model is deliberately stylized, thereby applicable across multiple cases and scales, to gain a deeper understanding more complex models might miss. The formal definitions of the three paradigms are designed to fit our mathematical framework (see below). Since we do not focus on intragenerational justice in this article, one agent suffices as a decision making subject, in contrast to a multiagent setting. We find that there exists no master paradigm between the three examined, i.e., a policy can be any combination of optimal or not, sustainable or not and safe or not. This is of special relevance to the climate system which may reside at the edge in the parameter regime where economic welfare optimization becomes neither sustainable nor safe. This suggests the use of more advanced paradigms to support decision making in climate policy.

Results

Stylized model of a human-environment tipping element. We use the mathematical framework of Markov Decision Processes[45,46], in which an agent makes decisions about how to interact with its environment (Fig. 1a). Our particular environment can reside in either a prosperous state, which provides immediate rewards (also called payoffs) to the agent, or a degraded state, from which the agent receives no payoff. At each time step, the agent chooses between two actions a, exerting either a high or low pressure on the environment. Depending on the current state s, the current action a and the subsequent state s', the agent receives an immediate reward r (Fig. 1b). At the prosperous state, taking the low pressure action the agent is guaranteed to receive reward r_l and remain at the prosperous state. However, taking the high pressure action, the agent may receive reward r_h (which is typically larger than r_l), but risks triggering a collapse of the environment to the degraded system state with non-zero probability δ and no immediate reward at all. From there, only the low pressure action opens the option to recover to the prosperous state with non-zero probability ρ.

For example, the high pressure action could correspond to emitting a business-as-usual amount of carbon to the atmosphere yielding a reward of high, short-term economic output as long as the system has not tipped. The low pressure action resembles emitting a reduced amount of carbon, assuming a lower short-term economic output for the guarantee to not trigger climate tipping elements into a disastrous state.

A policy π is a function that specifies what action a to apply at a system state s. The agent receives reward r_t at time step t. The value $v_\pi(s)$ of a state s under a given policy π is given by the expected value of the normalized accumulated discounted rewards r with discount factor $0 \le \gamma \le 1$ when starting in state $S_0 = s$ and following policy π:

$$v_\pi(s) = \mathbb{E}_\pi\left[\lim_{T\to\infty} \frac{\sum_{t=0}^{T} \gamma^t r_t}{\sum_{t=0}^{T} \gamma^t} | S_0 = s\right]. \tag{1}$$

Note that the discount factor actually denotes the farsightedness of the agent. Thus, $\gamma = 1$ corresponds to no discounting (weighting all rewards equally regardless of when they are

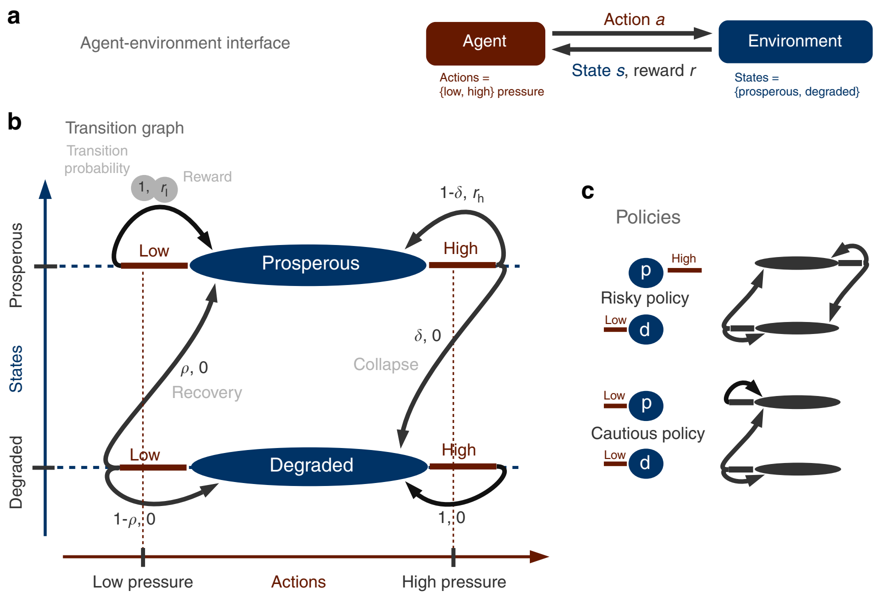


Fig. 1 Conceptual model of a human-environment tipping element. **a** Agent-environment interface: based on the state information and received reward, the agent chooses an action *a* from its actions set to gain rewards. **b** The transition graph gives state transition probabilities and corresponding rewards for all triples of state *s*, action *a*, next state *s′*, i.e., in state *s* the agent takes action *a* and moves to state *s′*. **c** Risky and cautious policies including the resulting Markov chains as a transition graph

expected), whereas $\gamma = 0$ corresponds to completely myopic, fully discounting agents.

Paradigm definitions. We classify policies according to whether they are economic welfare optimal or not, sustainable or not, and safe or not. For the sake of simplicity we focus on two deterministic policies, distinguishing whether the agent should apply the low or the high pressure action at the prosperous state (Fig. 1c): the risky policy ($\pi_r(p) = h$, $\pi_r(d) = l$), applying the high pressure action at the prosperous state and the low pressure one at the degraded state and the cautious policy ($\pi_c(p) = l$, $\pi_c(d) = l$), applying the low pressure action at the prosperous, as well as the degraded state.

A policy π is defined as optimal (in the economic welfare sense) if its value $v_\pi(s)$ (Eq. 1) for every state s is larger than or equal to the value of any other policy[46].

Based on the Brundtland Commission's report on sustainable development[26] a sustainable policy should fulfill two requirements: First, meet the needs of the present. We translate this formally into the agent evaluating the present state s as acceptable (similar to viable[17], tolerable[36] or desirable[47]), if its value (Eq. 1) exceeds a normatively chosen minimum acceptable value $r_{\min}$:

$$s \,\text{acceptable under}\, \pi \,\text{iff}\,\, v_\pi(s) \geq r_{\min} \tag{2}$$

Note, that the division of state space into acceptable and unacceptable states is not identical for all polices, but depends on the rewards receivable through executing a policy. Second, a sustainable policy should sustain the ability to meet the needs of the future[26].

We define a policy π as sustainable if every state the agent eventually visits under policy π is acceptable (Eq. 2).

Note that this reduction of sustainability to the one-dimensional value $v_\pi(s)$ has much similarity with the notion of weak sustainability[48].

The Safe Operating Space (SOS)[37] is typically defined as a subset of the whole state space $\mathcal{S}$, containing favorable system states bounded by thresholds[39,49]. In practice, the position of these potential tipping thresholds is always uncertain and the boundaries are placed at the lower end of the uncertainty zone. In that way the definition of the safe operating space states constitutes a normative judgment about the risk the decision maker is willing to tolerate. In the subsequent analyses we take the extreme position of no risk tolerance and identify the SOS with only the (more favorable) prosperous state, independent of the collapse probability δ.

We define a policy π as safe if every state the agents eventually visits under policy π lies within the SOS.

In contrast to acceptable and unacceptable states, safe states are independent of the policy used.

In summary, our stylized model of a human-environment tipping element depends on the five parameters δ, ρ, γ, r_l/r_h, $r_{\min}/r_h$: the probability of a collapse from the prosperous to the degraded state under the high pressure action δ, the probability of recovery from the degraded to the prosperous state under the low pressure action ρ, the agent's discount factor γ, the high reward receivable from the high pressure action when staying at the prosperous state r_h, the low reward receivable by taking the low pressure action at the prosperous state r_l, and the normatively chosen minimum acceptable reward $r_{\min}$ a state value must have to be perceived as acceptable under a certain policy. Since all three rewards come in arbitrary units, the policy classification only depends on their ratios.

Classification of risky and safe policy. Based on Eqs. 1 and 2 we analytically compute whether the risky and the cautious policy are

Reproduced from: W. Barfuss et al., When optimization for governing human-environment tipping elements is neither sustainable nor safe, Nat. Commun., vol. 9, no. 1, pp. 1-10, 2018, doi: 10.1038/s41467-018-04738-z. Published under Creative Commons Attribution License 4.0 (CC BY).

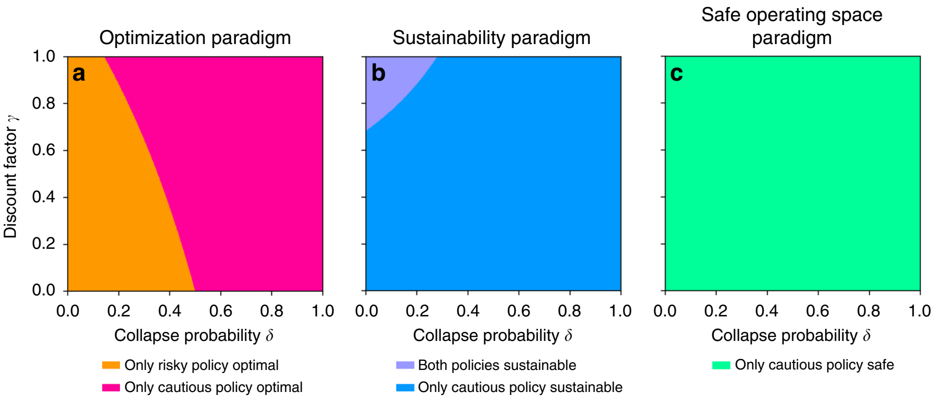


Fig. 2 Classification of the risky and cautious policy according to the three policy paradigms: **a** optimization, **b** sustainability, and **c** safe operating space in the model parameter space (shown here as collapse probability δ vs. discount factor γ); remaining parameters were chosen as $\rho = 0.2$, $r_l/r_h = 0.5$, $r_{min}/r_h = 0.3$ for illustration purposes. Colored regions result from analytically derived equations (see Methods). Depending on the parameter region, both risky and cautious policy can be optimal and sustainable. Only the cautious policy is safe

optimal or not, sustainable or not and safe or not depending on the model parameters (δ, ρ, γ, r_l/r_h, r_{min}/r_h) (see Methods and Fig. 2).

We observe that above a certain critical value of the collapse probability δ the cautious policy becomes optimal (Fig. 2a, pink), despite the smaller immediate reward $r_l = 0.5r_h$. This result confirms previous findings on optimal management with regime shifts[50].

Further, we find a decreasing critical collapse probability with increasing farsightedness γ. Hence, for more farsighted societies the risky policy is optimal only for small collapse probabilities δ (orange).

Provided the low pressure reward exceeds the normative minimum acceptable value threshold, $r_l \geq r_{min}$, then the cautious policy is sustainable for all parameter combinations δ, ρ, γ, r_l/r_h (Fig. 2b, blue and purple). Only for small collapse probabilities δ and simultaneously high farsightedness γ the risky policy becomes sustainable as well (purple). This is because in this parameter region the risky policy is acceptable also at the degraded state (Methods).

The cautious policy is a safe policy independently from the parameter combinations δ, ρ, γ, r_l/r_h, r_{min}/r_h (Fig. 2c, green). It is important to emphasize that there is no combination of parameters at which the risky policy is safe.

Relationships between paradigms. We find that policies can be classified along all logical combinations of the three examined paradigms (optimization, sustainability, safe operating space). This yields a classification of policies into eight different categories (Fig. 3).

In particular, optimal policies are not necessarily sustainable (opt and not sus: Fig. 3, red and yellow). This is the case if the normative value threshold r_{min} is too large. The cautious policy does not return enough value to be sustainable ($r_l < r_{min}$, yellow) and the risky policy at the degraded state produces too little future reward to be sustainable, due to the low chance of recovery and lack of farsightedness.

Nor are optimal policies necessarily safe (opt and not safe: Fig. 3, red and purple). This occurs in parameter regions where the risky policy is optimal. The risky policy cannot be safe because of the risk of collapse to the degraded state.

A safe policy does not necessarily imply a sustainable policy either (safe and not sus: Fig. 3, green and yellow). When the normative threshold value for sustainability r_{min} exceeds the reward from a low pressure action r_l: $r_{min} > r_l$, then the cautious policy is safe but not sustainable. Following a similar line of argument, the SOS concept[37] has been extended to a Safe And Just Operating Space (SAJOS) which additionally accounts for social indicators[51], such as the number of people living in extreme poverty. Thus, SAJOS policies can be interpreted as the overlap of safe with sustainable policies. Within our model, we can give a definite criterion for when this form of SAJOS exists: as long as the reward from a low pressure action r_l exceeds the normative threshold value r_{min} ($r_l > r_{min}$), the cautious policy is both safe and sustainable (Fig. 3, cyan and gray).

However, there exist also sustainable policies outside the SOS (sus and not safe: Fig. 3, blue and purple.) These are risky policies (hence, not safe) with simultaneously high farsightedness γ and low collapse probability δ. At those parameter regions the degraded state is still evaluated as acceptable due to sufficient anticipated future rewards and therefore the risky policy is sustainable. The circumstance that parameter regimes exist that are sustainable but not safe and vice versa clearly stems from our definition of sustainability which resembles a form of weak sustainability[48]. By doing so we can conceptually separate issues of environmentally safe and socially just without compromising the target of a safe and just parameter space regime.

Note that this classification into the eight different policy paradigm combinations also applies to the case of absolute farsightedness ($\gamma = 1$; see the tops of Fig. 3b–e). Thus, the trade-offs between the examined paradigms do not vanish, as one might presume considering the debate about appropriate discount rates[14,52].

Volume of paradigm combinations. So far, we have visualized the parameter space of our stylized tipping element model in two dimensional sections and fixed the remaining parameters for illustrative purposes. By doing so, we showed the mutual dependence between parameters, foremost the discount factor γ and the collapse probability δ. However, in the light of considerable parameter uncertainty we ask how large the eight regimes of paradigm combinations are, given the whole parameter space (Fig. 4).

We observe the most likely option to be the regime that is neither optimal, neither sustainable nor safe followed by the parameter sweet spot regime in which all paradigms yield the

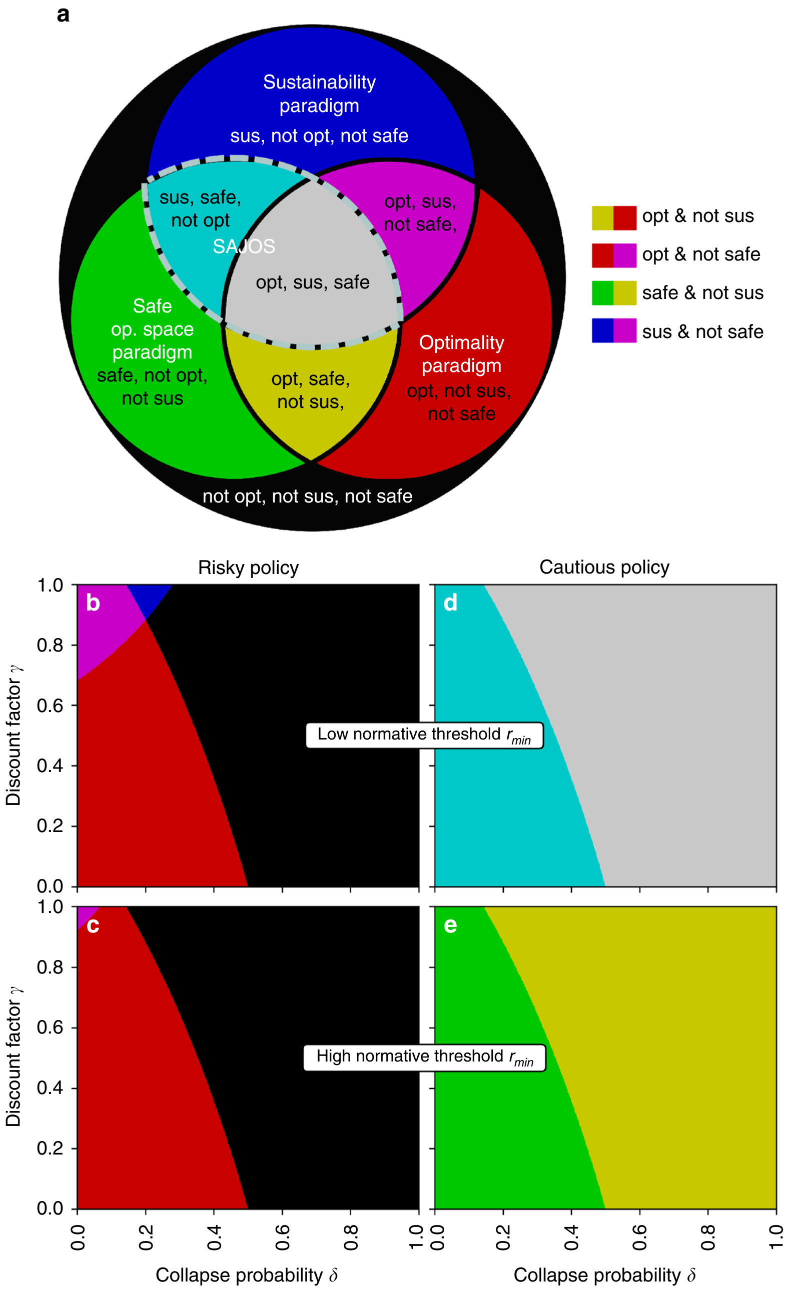


Fig. 3 Paradigms classification for risky and cautious policy. There exist policies in parameter space of our model for all logical combinations of paradigm classifications. **a** i.e. a policy can be any combination of (not) optimal, (not) sustainable and (not) safe. Remaining parameters where chosen as $\rho = 0.2$, $r_l/r_h = 0.5$ for illustration purposes (c.f. Methods). For a sufficiently low normative threshold value $r_{min} \le r_l$ (here $r_{min}/r_h = 0.3$) a Safe And Just Operating Space (SAJOS) exists, which we identified as the overlap of safe and sustainable policies **b**, **d** (gray and cyan area). For a sufficiently large $r_{min} > r_l$ (here $r_{min}/r_h = 0.7$) a SAJOS does not exist **c**, **e**

cautious policy as optimal, sustainable and safe. Together they constitute a parameter space volume of approx. 45% in which the three paradigms of economic optimization, sustainability and safe operating space align with each other in yielding the same policy. Interestingly, the third likeliest option is the paradigm combination in which the risky policy is optimal but neither sustainable nor safe. This is the most likeliest parameter regime among those where the paradigms yield different policies. Thus, blindly applying economic optimization in a our stylized tipping element has a significant chance of leading to policies that are neither sustainable nor safe.

On the other hand, the volume of the safe and just operating space (gray and cyan bars in Fig. 4) is comparable to the most likeliest (black) regime. Thus, about one out of four random

Reproduced from: W. Barfuss et al., When optimization for governing human-environment tipping elements is neither sustainable nor safe, Nat. Commun., vol. 9, no. 1, pp. 1-10, 2018, doi: 10.1038/s41467-018-04738-z. Published under Creative Commons Attribution License 4.0 (CC BY).

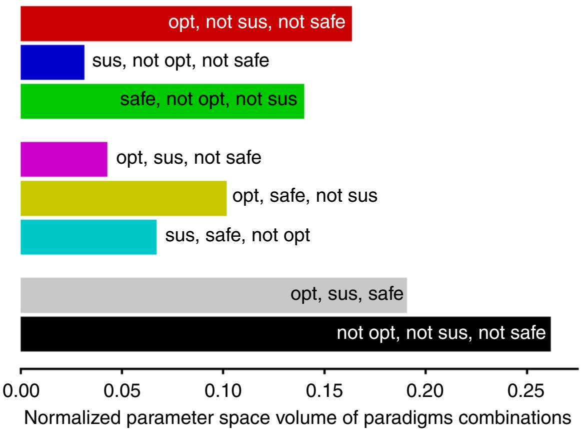


Fig. 4 Ratios of parameter space volumes for all eight paradigms combination. All parameters (δ, ρ, γ, r_l/r_h, r_{min}/r_h) were chosen linearly between 0 and 1 for both the risky and the cautious policy. As a direct consequence of our definitions of the safe operating space paradigm and the cautious and risky policy, all paradigm combinations which are safe correspond to the use of the cautious policy, in all others the risky policy was applied. A random decision making agent within a random tipping element will most likely end up with a policy that is neither optimal, neither sustainable nor safe, followed by the parameter sweet spot regime where the policy is simultaneously optimal, sustainable and safe. Interestingly, the third likeliest option is a parameter regime which is optimal, but neither sustainable nor safe

decision making agents interacting with a random tipping element will end up in the safe and just operating space.

Application to real-world human-environment tipping elements. The above policy classification offers valuable insights for the governance of real-world human-environment systems. We discuss how our analysis relates to the cases of the climate system, fisheries and farming. Our purpose is to gain a qualitative understanding how our model relates to important real-world challenges of environmental governance, not a detailed assessment of the latter. Therefore, we roughly estimate the respective collapse and recovery probabilities per time step δ and ρ of our model via the typical timescales on which these systems remain in one state or the other (see Methods). Additionally, we added a parameter sensitivity analysis by visualizing the likelihood of ending up in a certain parameter regime by color gradients between regimes (Fig. 5).

Regarding the climate system, we acknowledge that several interacting tipping elements contribute to the system's behavior[2] and its representation as a single tipping element is a huge simplification on its own. Nevertheless, we assume that the current state of the climate system is still comparable to the prosperous one of our model and relevant timescales for triggering a collapse of 30 to 50 years under business-as-usual socio-economic development scenarios[2,53,54]. Regarding the recovery timescale it has been shown that human perturbations of the climate system already changed its trajectory on a multi-millennial timescale[55,56]. Therefore we assume a recovery probability per time step ρ close to zero (Fig. 5).

For sufficiently large collapse probabilities (collapse time scale near 20 years and smaller), the climate system is likely to reside in a parameter sweet spot (gray area), where applying an optimization, sustainability or SOS paradigm results in the cautious policy as the advisable way of governing the climate system. However, if the collapse probability per time step is smaller (collapse time scale near 50 years and larger) the situation is different. Here, an SOS and a sustainable paradigm would still yield the cautious policy (Fig. 5, cyan), but an optimization paradigm is likely to give the risky policy (Fig. 5, red), which at this point is neither sustainable nor safe. We conclude that in climate policy, economic welfare optimization alone may neither be sustainable nor safe.

For fishery systems, both transition probabilities certainly depend on a variety of factors, e.g., fisher's technical and cultural traits or the dominant fish species in the system, as well as external factors such as climate change influencing habitat condition[57,58]. The timescale of a fisheries collapse has been shown to lie within decades[59]. Roughly consistent with observational and modeled data from the Baltic sea, where the stable regime of high cod biomass lasted approximately from 1970 to 1990[57,60], we assume a typical collapse timescale of around 20 years. Concerning the typical recovery time scale, successful attempts of fish stocks recovery lasted for decades[61], but is estimated to generally exceed this duration[62]. We therefore assume a larger typical recovery timescale of around 50 years. The color gradient in Fig. 5 at the fisheries point does not clearly single out a paradigms regime, indicating the dependence on the other parameters at this point. A risky policy might be economically optimal (Fig. 5, red), but leads eventually to the collapse of fish stock (c.f[59].). At the collapsed and degraded state the conditions for the fishers are not acceptable. Therefore they have to leave the system and cannot wait for the fish's recovery. But further investigation is needed to reduce the uncertainty with respect to the other parameters.

Last, we look at the case of land degradation by farming in our stylized model. Land degradation and restoration is a complex topic with many influencing factors[63]. Nevertheless, land degradation by farming has been identified as a tipping element by Kinzig and others[64], where the authors discuss the case of the western Australian wheatbelt with a typical collapse timescale of about 100 years. Soil recovery is estimated to take place within 20 to 1000 years[65], which is roughly consistent to Kitzing et al., where the duration to reach equilibrium again is estimated with up to 300 years. We therefore assume a typical recovery timescale of about 300 years. In contrast to climate and fisheries, the transition probabilities we associated with the process of land degradation by farming suggest, that here an optimality paradigm is very likely to yield the risky policy which is neither sustainable nor safe despite considerate parameter uncertainty (red area in Fig. 5).

Taken together, it is interesting to see that in particular the climate system may reside at the edge of the parameter regime where economic welfare optimization becomes neither sustainable nor safe (Fig. 3). For land degradation by farming, our assessment suggests that an optimal policy is likely to yield a non-sustainable and non-safe policy whereas for fisheries the situation is less clear.

Discussion

Overall, our results show that there exists no master paradigm among the three examined in our model of environmental governance of a stylized tipping element. Policies can be classified by any combination of optimal, sustainable and safe. A master paradigm, in contrast, would guarantee fulfilling requirements imposed by other paradigms. Consequently, the selection of appropriate policy paradigms, especially in more complex settings and models, can be critical for effective environmental governance.

Specifically, our results show theoretically, as well as empirically that economic welfare optimization for managing tipping

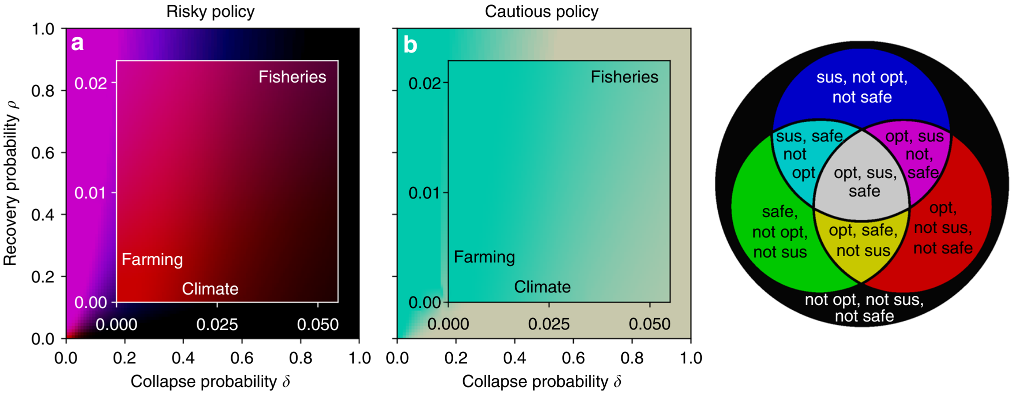


Fig. 5 Human-environment systems in paradigms classification. For risky (**a**) and (**b**) cautious policy here shown in model parameter space of collapse probability δ versus recovery probability ρ. Color indicates the paradigms combination similarly as in Fig. 3. Here, additional gradual changes between the color regimes indicate the probability of being in a certain paradigms combinations regime under parameter uncertainty ranges. Remaining parameters where chosen linearly within the range of $0.95 \le \gamma \le 0.99$, $0.3 \le r_l/r_h \le 0.7$, $0.1 \le r_{min}/r_h \le 0.5$. The approx. transition probabilities δ and ρ were assigned to the human-environment systems climate, fisheries and farming agriculture according to the timescale of the average time spent in one state (see Methods). For farming, a risky policy is likely to be optimal but neither sustainable nor safe. The parameter uncertainty of the other parameters does not allow a clear statement in which parameter regime fisheries are likely to fall. The climate system may lie at the edge of the sweet spot, where all paradigms yield the cautious policy. However, for smaller collapse probability δ optimization is more likely to yield the risky policy, which becomes also neither sustainable nor safe at this point. This suggests the use of other paradigms for climate policy making

elements may be neither sustainable nor safe. For example, the volume of the corresponding paradigm combination in parameter space is the largest among those in which the three paradigms actually yield different policies. This suggests the conclusion that the mere structure of a tipping element causes a comparable high chance of obtaining a policy that is neither sustainable nor safe when blindly following an optimization paradigm. On the other hand, our model also indicates parameter regimes where economic optimization can safely and sustainably be used.

We derived simple heuristics to anticipate when a policy is economic welfare optimal, sustainable and safe. A risky policy may be optimal when the probability of collapse and/or the farsightedness are sufficiently small. It may be sustainable when the probability of a collapse is sufficiently small but the farsightedness is sufficiently large. However, it cannot be safe. A cautious policy may be optimal when the collapse probability and/or the farsightedness are sufficiently large. It is sustainable if its immediate reward exceeds the normatively chosen minimum acceptable reward and it is always safe. The absence of a master paradigm is of special relevance for governing the climate system, since the latter may reside at the edge between parameter regimes where economic welfare optimization becomes neither sustainable nor safe.

Extensions are possible in many directions. Constrained optimization[24] is a straight-forward way to combine the paradigms examined. Policy makers could aim for the maximum economic welfare delivering a policy that is safe and sustainable, or least-cost safe target strategies[15]. This is certainly a better approach than relying on economic welfare optimization alone for model-based policy advice. Examples of models for policy advice certainly include integrated assessment models or the use of the maximum sustainable yield in fisheries management. However, one might not desire to obtain the welfare optimal safe and sustainable policy but e.g., the most resilient one, which calls for an operationalization of modern social-ecological resilience concepts[66].

The application of our model to real-world systems in this article is of qualitative, illustrative nature. A more detailed analysis of real world tipping elements in which typical transition probabilities might be estimated from empirical time series could be a way forward to systematize and draw lessons from the multitude of human-environmental tipping elements[67].

Applying our analyses to larger, more complex Markov decision processes would be a way to extend the understanding of the relationships between the paradigms examined. Moreover, it may be desirable to include further policy paradigms into the analyses, e.g., aiming for a large option space of future decision makers[30,68]. Based on such analyses, policy makers could make better informed decisions on how to translate the Sustainable Development Goals and the Paris climate agreement into concrete policy implementations.

Methods

Derivation of value functions. There are four deterministic policies in our Markov decision process model: (1) $\pi_r(p) = h$, $\pi_r(d) = l$, (2) $\pi_c(p) = l$, $\pi_c(d) = l$, (3) $\pi_3(p) = h$, $\pi_3(d) = h$, (4) $\pi_4(p) = l$, $\pi_4(d) = h$. We concentrate on deterministic policies only to simplify the calculation without loss of generality, because if an optimal policy exits there exits also a deterministic optimal policy[46]. We further focus here only on the first two policies, named the risky and the cautious policy, since the remaining two apply a high pressure action at the degraded state. This will trap the agent at this position for eternity without receiving any reward. The math on these policies is left to the interested reader.

In the following we derive the analytical expressions of the state values of these policies as functions of the parameters (δ, ρ, γ, r_l, r_h). From Eq. 1 and for $\gamma < 1$ one can derive the recursive relationship between state values, known as the Bellman Equation[69]:

$$v_\pi(s) = \sum_{s'} p(s'|s, \pi(s))[(1-\gamma)r(s, \pi(s), s') + \gamma v_\pi(s')] \tag{3}$$

with $p(s'|s, \pi(s))$ being the probability to enter state s' given the agent has started in state s and applied action $\pi(s)$.

Applied to our model the value for the prosperous state reads

$$v_\pi(p) = \begin{cases} \delta\gamma v_\pi(d) + (1-\delta)[(1-\gamma)r_h + \gamma v_\pi(p)] & \text{for } a = h \\ (1-\gamma)r_l + \gamma v_\pi(p) & \text{for } a = l \end{cases}. \tag{4}$$

The value for the degraded state is given by

$$v_\pi(d) = \begin{cases} \gamma v_\pi(d) & \text{for } a = h \\ (1-\rho)\gamma v_\pi(d) + \rho\gamma v_\pi(p) & \text{for } a = l \end{cases}. \tag{5}$$

To obtain the explicit state values for the risky policy ($\pi_r(p) = h$, $\pi_r(d) = l$) we solve the system of equations

$$v_{\pi_r}(p) = \delta\gamma v_{\pi_r}(d) + (1-\delta)\Big[(1-\gamma)r_\mathrm{h} + \gamma v_{\pi_r}(p)\Big] \tag{6}$$

$$v_{\pi_r}(d) = (1-\rho)\gamma v_{\pi_r}(d) + \rho\gamma v_{\pi_r}(p), \tag{7}$$

which yields

$$v_{\pi_r}(p) = r_\mathrm{h}\frac{(1-\delta)(1-(1-\rho)\gamma)}{1-(1-\delta-\rho)\gamma} \tag{8}$$

$$v_{\pi_r}(d) = r_\mathrm{h}\frac{(1-\delta)\rho\gamma}{1-(1-\delta-\rho)\gamma}. \tag{9}$$

To obtain the explicit state values for the cautious policy ($\pi_c(p) = l$, $\pi_c(d) = l$) we solve the system of equations

$$v_{\pi_c}(p) = (1-\gamma)r_\mathrm{l} + \gamma v_{\pi_c}(p) \tag{10}$$

$$v_{\pi_c}(d) = (1-\rho)\gamma v_{\pi_c}(d) + \rho\gamma v_{\pi_c}(p), \tag{11}$$

which yields

$$v_{\pi_c}(p) = r_\mathrm{l} \tag{12}$$

$$v_{\pi_c}(d) = \frac{\rho\gamma r_\mathrm{l}}{1-(1-\rho)\gamma}. \tag{13}$$

For $\gamma = 1$ we compute the values v_π (which are independent from the initial state for $\gamma = 1$) by multiplying the stationary state of the effective Markov chain with the reward vector $\mathbf{r}^\pi \in \mathbb{R}^{|\mathcal{S}|}$ whose components read

$$r^\pi_s = \sum_{s'} p(s'|s,\ \pi(s))r(s, \pi(s), s'). \tag{14}$$

The components of the transition matrix $\mathbf{P}^\pi$ of the effective Markov chain read

$$P^\pi_{s's} = p(s'|\pi(s), s). \tag{15}$$

The stationary state $\boldsymbol{\sigma}_\pi$ is the normalized eigenvector of the transition matrix with eigenvalue one. Hence,

$$v_\pi = \sigma_\pi \cdot \mathbf{r}^\pi. \tag{16}$$

Performing this calculation for risky and cautious policy explicitly yields consistent results with the calculation for $0 \le \gamma < 1$ from above. For $\gamma = 1$ the value v_π can be obtained by simply inserting $\gamma = 1$ into Eqs. 8 and 9 for the risky policy and Eqs. 12 and 13 for the cautious policy.

Analytical expressions for paradigm policy classification. To derive the analytical expression of the hypersurface in parameter space that separates the regions where either the risky or the cautious policy is optimal we set $v_{\pi_r}(p) \overset{\text{set}}{=} v_{\pi_c}(p)$ (or equivalently $v_{\pi_r}(d) \overset{\text{set}}{=} v_{\pi_c}(d)$, since the parameter combination where a policy is optimal is independent from the state) and implicitly obtain

$$\tilde{r}_\mathrm{h} \cdot \left(1-\tilde{\delta}\right)(1-\tilde{\gamma}(1-\tilde{\rho})) = \tilde{r}_\mathrm{l} \cdot \left(1-\tilde{\gamma}\left(1-\tilde{\delta}-\tilde{\rho}\right)\right). \tag{17}$$

To obtain the hypersurface that separates state s being acceptable from being not acceptable under policy π we apply the definition from Eq. 2: $v_\pi(s) \overset{\text{set}}{=} r_\mathrm{min}$. Hence, for the risky policy at the prosperous state we set $v_{\pi_r}(p) \overset{\text{set}}{=} r_\mathrm{min}$ and obtain implicitly

$$\tilde{r}_\mathrm{h} \cdot \left(1-\tilde{\delta}\right)(1-\tilde{\gamma}(1-\tilde{\rho})) = \tilde{r}_\mathrm{min} \cdot \left(1-\tilde{\gamma}\left(1-\tilde{\delta}-\tilde{\rho}\right)\right). \tag{18}$$

For the risky policy at the degraded state we set $v_{\pi_r}(d) \overset{\text{set}}{=} r_\mathrm{min}$ and obtain implicitly

$$\tilde{r}_\mathrm{h} \cdot \left(1-\tilde{\delta}\right)\tilde{\rho}\tilde{\gamma} = \tilde{r}_\mathrm{min} \cdot \left(1-\tilde{\gamma}\left(1-\tilde{\delta}-\tilde{\rho}\right)\right). \tag{19}$$

For the cautious policy at the prosperous state we set $v_{\pi_c}(p) \overset{\text{set}}{=} r_\mathrm{min}$ and obtain implicitly

$$\tilde{r}_\mathrm{l} = \tilde{r}_\mathrm{min}. \tag{20}$$

For the cautious policy at the degraded state we set $v_{\pi_c}(d) \overset{\text{set}}{=} r_\mathrm{min}$ and obtain implicitly

$$\tilde{r}_\mathrm{l} \cdot \tilde{\rho}\tilde{\gamma} = \tilde{r}_\mathrm{min} \cdot (1-\tilde{\gamma}(1-\tilde{\rho})) \tag{21}$$

To get from acceptability to sustainability for the risky policy one has to logically combine Eqs. 18 and 19. The risky policy is sustainable only if both the prosperous and the degraded state are acceptable since it will visit both states recurrently. The safe policy is sustainable exactly where the prosperous state is acceptable since it will eventually end up and remain at the prosperous state. Supplementary Fig. 1 shows an example of the acceptability division of state-parameter space and the resulting sustainability division.

The division of the parameter space according the safe operating space paradigm is obvious from its definition. Only the cautious policy is a safe policy since it will eventually end up and remain in the prosperous, safe operating space state. The risky policy switches recurrently between the prosperous and the degraded which makes it, by definition, not safe.

Conversion of timescales to transition probabilities. Let p be the probability per time step that a system state will transition into another state. The average number of time steps the system will be in that state is given by $\langle \mathrm{N} \rangle = (1-p)/p$. Inverting yields $p = 1/(\langle N \rangle + 1)$. We map a model time step to a year. Thus, a collapse time scale of e.g., 50 years corresponds to a collapse probability of $\delta \approx 0.02$. Supplementary Tab. 1 shows the assumed transition timescales and corresponding transition probabilities.

Code availability. Python code for the reproduction of the reported results plus interactive versions of the figures can be downloaded from https://github.com/wbarfuss/Paradigms.

Data availability. Data sharing not applicable to this article as no datasets were stored on disk during the production of the figures (see Code availability).

Received: 2 February 2018 Accepted: 16 May 2018
Published online: 15 June 2018

References

1. Griggs, D. et al. Policy: sustainable development goals for people and planet. *Nature* **495**, 305–307 (2013).
2. Lenton, T. M. et al. Tipping elements in the Earth's climate system. *Proc. Natl Acad. Sci.* **105**, 1786–1793 (2008).
3. Schellnhuber, H. J. Tipping elements in the earth system. *Proc. Natl Acad. Sci.* **106**, 20561–20563 (2009).
4. Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., Walker, B. Catastrophic shifts in ecosystems. *Nature* **413**, 591–596 (2001).
5. Lade, S. J., Tavoni, A., Levin, S. A. & Schlüter, M. Regime shifts in a social-ecological system. *Theor. Ecol.* **6**, 359–372 (2013).
6. Donges, J. F. et al. Closing the loop: reconnecting human dynamics to earth system science. *Anthr. Rev.* **4**, 151–157 (2017).
7. Anderies, J. M., Rodriguez, A. A., Janssen, M. A. & Cifdaloz, O. Panaceas, uncertainty, and the robust control framework in sustainability science. *Proc. Natl Acad. Sci.* **104**, 15194–15199 (2007).
8. Polasky, S., Carpenter, S. R., Folke, C. & Keeler, B. Decision-making under great uncertainty: environmental management in an era of global change. *Trends Ecol. Evol.* **26**, 398–404 (2011).
9. Irwin, E. G., Gopalakrishnan, S. & Randall, A. Welfare, wealth, and sustainability. *Annu. Rev. Resour. Econ.* **8**, 77–98 (2016).
10. Farmer, J. D., Hepburn, C., Mealy, P. & Teytelboym, A. A third wave in the economics of climate change. *Environ. Resour. Econ.* **62**, 329–357 (2015).
11. Crépin, A.-S., Biggs, R., Polasky, S., Troell, M. & de Zeeuw, A. Regime shifts and management. *Ecol. Econ.* **84**, 15–22 (2012).

12. Perman, R., Ma, Y., McGilvray, J. & Common, M. *Natural resource and environmental economics*. (Pearson Education, Essex, 2003).
13. Weyant, J. Integrated assessment of climate change: state of the literature. *J. Benefit-Cost. Anal.* **5**, 377–409 (2014).
14. Stern, N. The economics of climate change. *Am. Econ. Rev.* **98**, 1–37 (2008).
15. Ackerman, F., DeCanio, S. J., Howarth, R. B. & Sheeran, K. Limitations of integrated assessment models of climate change. *Clim. Change* **95**, 297–315 (2009).
16. Woodward, R. T. & Tomberlin, D. Practical precautionary resource management using robust optimization. *Environ. Manag.* **54**, 828–839 (2014).
17. Martinet, V. & Doyen, L. Sustainability of an economy with an exhaustible resource: a viable control approach. *Resour. Energy Econ.* **29**, 17–39 (2007).
18. De Lara, M. & Doyen, L. *Sustainable Management of Natural Resources: Mathematical Models and Methods*. (Springer Science & Business Media, 2008).
19. Rougé, C., Mathias, J. -D. & Deffuant, G. Extending the viability theory framework of resilience to uncertain dynamics, and application to lake eutrophication. *Ecol. Indic.* **29**, 420–433 (2013).
20. Chadès, I., et al. Optimization methods to solve adaptive management problems. *Theoretical Ecology*, 1–20 (2017).
21. Branke, J., Deb, K., Miettinen, K., Słowinski, R. *Multi-objective Optimization: Interactive and Evolutionary Approaches*. (Springer-Verlag Berlin Heidelberg, 2008).
22. Greco, S., Ehrgott, M. & Figueira, J. R. *Multiple Criteria Decision Analysis*. (Springer Science+Business Media, New York, 2005).
23. Ehrgott, M. *Multicriteria Optimization*. (Springer Science & Business Media 2006).
24. Altman, E. *Constrained Markov Decision Processes*, Vol. 7 (CRC Press, 1999).
25. Meadows, D. H., Goldsmith, E. & Meadows, P. *The Limits of Growth*, Vol. 381 (Earth Island Limited, London, 1972).
26. World Commission on Environment and Development. *Our Common Future*. Technical report (1987).
27. Pezzey, J. Sustainable development concepts. *World Bank Environ. Pap.* **1**, 45 (1992).
28. Pezzey, J. C. V. Sustainability Constraints versus "Optimality" versus Intertemporal Concern, and Axioms versus Data. *Land Econ.* **73**, 448–466 (1997).
29. Arrow, K. J., Dasgupta, P., Goulder, L. H., Mumford, K. J. & Oleson, K. Sustainability and the measurement of wealth. *Environ. Dev. Econ.* **17**, 317–353 (2012).
30. Fleurbaey, M. On sustainability and social welfare. *J. Environ. Econ. Manag.* **71**, 34–53 (2015).
31. Gerlagh, R. Generous sustainability. *Ecol. Econ.* **136**, 94–100 (2017).
32. Pezzey, J. C. V. One-sided sustainability tests with amenities, and changes in technology, trade and population. *J. Environ. Econ. Manag.* **48**, 613–631 (2004).
33. Dasgupta, P. & Karl-Göran, M. *The Economics of Non-convex Ecosystems*, Vol. 4. (Springer Science & Business Media 2006).
34. Lontzek, T. S., Cai, Y., Judd, K. L. & Lenton, T. M. Stochastic integrated assessment of climate tipping points indicates the need for strict climate policy. *Nat. Clim. Change* **5**, 441 (2015).
35. Cai, Y., Lenton, T. M. & Lontzek, T. S. Risk of multiple interacting tipping points should encourage rapid co 2 emission reduction. *Nat. Clim. Change* **6**, 520 (2016).
36. Petschel-Held, Gerhard, Schellnhuber, Hans-Joachim, Bruckner, Thomas, Toth, FerencL. & Hasselmann, Klaus The tolerable windows approach: theoretical and methodological foundations. *Clim. Change* **41**, 303–331 (1999).
37. Rockström, J. et al. A safe operating space for humanity. *Nature* **461**, 472–475 (2009).
38. Dearing, J. A. et al. Safe and just operating spaces for regional social-ecological systems. *Glob. Environ. Change* **28**, 227–238 (2014).
39. Carpenter, S. R., Brock, W. A., Folke, C., van Nes, E. H. & Scheffer, M. Allowing variance may enlarge the safe operating space for exploited ecosystems. *Proc. Natl Acad. Sci.* **112**, 14384–14389 (2015).
40. Folke, C. et al. Resilience thinking: integrating resilience, adaptability and transformability. *Ecol. Soc.* **15**, 20 (2010).
41. Raffensperger, C. & Tickner, J. A. *Protecting Public Health and the Environment: Implementing the Precautionary Principle*. (Island Press, Wahington, DC, 1999).
42. Fischer, J. et al. Integrating resilience thinking and optimisation for conservation. *Trends Ecol. Evol.* **24**, 549–554 (2009).
43. Karl-Göran, M. & Li, C.-Z. Measuring sustainability under regime shift uncertainty: a resilience pricing approach. *Environ. Dev. Econ.* **15**, 707–719 (2010).
44. Derissen, S., Quaas, M. F. & Baumgärtner, S. The relationship between resilience and sustainability of ecological-economic systems. *Ecol. Econ.* **70**, 1121–1128 (2011).
45. Bellman, R. A Markovian decision process. *Indiana Univ. Math. J.* **6**, 679–684 (1957).
46. Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. (John Wiley and Sons, Inc, Hoboken, New Jersey, 2005).
47. Heitzig, J., Kittel, T., Donges, J. F. & Molkenthin, N. Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the Earth system. *Earth Syst. Dyn.* **7**, 21–50 (2016).
48. Neumayer, E. *Weak Versus Strong Sustainability: Exploring the Limits of Two Opposing Paradigms*. (Edward Elgar Publishing, 2003).
49. Steffen, W. et al. Planetary boundaries: guiding human development on a changing planet. *Science* **347**, 1259855 (2015).
50. Polasky, S., Zeeuw, A. D. & Wagener, F. Optimal management with potential regime shifts. *J. Environ. Econ. Manag.* **62**, 229–240 (2011).
51. Raworth, K. A doughnut for the anthropocene: humanity's compass in the 21st century. *Lancet Planet. Health* **1**, e48–e49 (2017).
52. Nordhaus, W. D. A review of the Stern review on the economics of climate change. *J. Econ. Lit.* **45**, 686–702 (2007).
53. Schellnhuber, H. J., Rahmstorf, S. & Winkelmann, R. Why the right climate target was agreed in paris. *Nat. Clim. Change* **6**, 649–653 (2016).
54. Rockström, J. et al. A roadmap for rapid decarbonization. *Science* **355**, 1269–1271 (2017).
55. Clark, P. U. et al. Consequences of twenty-first-century policy for multi-millennial climate and sea-level change. *Nat. Clim. Change* **6**, 360 (2016).
56. Ganopolski, A., Winkelmann, R. & Schellnhuber, H. J. Critical insolation–co2 relation for diagnosing past and future glacial inception. *Nature* **529**, 200–203 (2016).
57. Moellmann, C. et al. Reorganization of a large marine ecosystem due to atmospheric and anthropogenic pressure: a discontinuous regime shift in the central baltic sea. *Glob. Change Biol.* **15**, 1377–1393 (2009).
58. Worm, B. et al. Rebuilding global fisheries. *Science* **325**, 578–585 (2009).
59. Costello, C., Gaines, S. D. & Lynham, J. Can catch shares prevent fisheries collapse? *Science* **321**, 1678–1681 (2008).
60. Österblom, H. et al. Human-induced trophic cascades and ecological regime shifts in the baltic sea. *Ecosystems* **10**, 877–889 (2007).
61. Hutchings, J. A. & Reynolds, J. D. Marine fish population collapses: consequences for recovery and extinction risk. *AIBS Bull.* **54**, 297–309 (2004).
62. Caddy, J. F. & Agnew, D. J. An overview of recent global experience with recovery plans for depleted marine resources and suggested guidelines for recovery planning. *Rev. Fish. Biol. Fish.* **14**, 43 (2004).
63. Blaikie, P. & Brookfield, H. *Land Degradation and Society*. (Routledge, 2015).
64. Kinzig, A.P., et al. Resilience and regime shifts: assessing cascading effects. *Ecol. Soc.* **11**, 20 (2006).
65. Horrigan, L., Lawrence, R. S. & Walker, P. How sustainable agriculture can address the environmental and human health harms of industrial agriculture. *Environ. Health Perspect.* **110**, 445 (2002).
66. Donges, J. F. & Barfuss, W. From math to metaphors and back again: social-ecological resilience from a multi-agent-environment perspective. *GAIA-Ecol. Perspect. Sci. Soc.* **26**, 182–190 (2017).
67. Rocha, J., Yletyinen, J., Biggs, R., Blenckner, T. & Peterson, G. Marine regime shifts: drivers and impacts on ecosystems services. *Philos. Trans. R. Soc. B* **370**, 20130273 (2015).
68. Schellnhuber, H. -J. Earth system analysis and the second Copernican revolution. *Nature* **402**, C19–C23 (1999).
69. Bellman, R. *Dynamic Programming*. (Princeton University Press, 1957).

Acknowledgements

This work was developed in the context of the COPAN project on Coevolutionary Pathways in the Earth system at the Potsdam Institute for Climate Impact Research. The authors are grateful for financial support from the Heinrich-Böll-Foundation, the Stordalen Foundation (via the Planetary Boundaries Research Network PB.net), the Earth League's EarthDoc program, the Leibniz Association (project DOMINOES) and the Swedish Research Council Formas (Project Grant 2014-589). We thank David Collste, Jobst Heitzig, Antoine Levesque, Finn Müller-Hansen and Maja Schlüter for discussions and comments on the manuscript.

Author contributions

W.B. designed and analyzed the model with assistance from J.F.D. and S.L. J.F.D and J.K. supervised the project. All authors wrote the manuscript.

Reproduced from: W. Barfuss et al., When optimization for governing human-environment tipping elements is neither sustainable nor safe, Nat. Commun., vol. 9, no. 1, pp. 1-10, 2018, doi: 10.1038/s41467-018-04738-z. Published under Creative Commons Attribution License 4.0 (CC BY).

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-04738-z.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reproduced from: W. Barfuss et al., When optimization for governing human-environment tipping elements is neither sustainable nor safe, Nat. Commun., vol. 9, no. 1, pp. 1-10, 2018, doi: 10.1038/s41467-018-04738-z. Published under Creative Commons Attribution License 4.0 (CC BY).

sustainability

MDPI

Article

A Thought Experiment on Sustainable Management of the Earth System

Jobst Heitzig [1,*], Wolfram Barfuss [1,2] and Jonathan F. Donges [1,3]

[1] Potsdam Institute for Climate Impact Research, P.O. Box 60 12 03, 14412 Potsdam, Germany; barfuss@pik-potsdam.de (W.B.); donges@pik-potsdam.de (J.F.D.)
[2] Department of Physics, Humboldt-Universität zu Berlin, 12489 Berlin, Germany
[3] Stockholm Resilience Centre, Stockholm University, SE-10691 Stockholm, Sweden
[*] Correspondence: heitzig@pik-potsdam.de; Tel.: +49-331-288-2692

Received: 30 April 2018; Accepted: 7 June 2018; Published: 11 June 2018

Abstract: We introduce and analyze a simple formal thought experiment designed to reflect a qualitative decision dilemma humanity might currently face in view of anthropogenic climate change. In this exercise, each generation can choose between two options, either setting humanity on a pathway to certain high wellbeing after one generation of suffering, or leaving the next generation in the same state as the current one with the same options, but facing a continuous risk of permanent collapse. We analyze this abstract setup regarding the question of what the right choice would be both in a rationality-based framework including optimal control, welfare economics, and game theory, and by means of other approaches based on the notions of responsibility, safe operating spaces, and sustainability paradigms. Across these different approaches, we confirm the intuition that a focus on the long-term future makes the first option more attractive while a focus on equality across generations favors the second. Despite this, we generally find a large diversity and disagreement of assessments both between and within these different approaches, suggesting a strong dependence on the choice of the normative framework used. This implies that policy measures selected to achieve targets such as the United Nations Sustainable Development Goals can depend strongly on the normative framework applied and specific care needs to be taken with regard to the choice of such frameworks.

Keywords: decision dilemma; intergenerational welfare; time horizon; risk attitude; inequality aversion; fairness; responsibility; sustainability paradigms

1. Introduction

The growing debate about concepts such as the Anthropocene [1], Planetary Boundaries [2–4], and Safe and Just Operating Spaces for Humanity [5], and the evidence about climate change and approaching tipping elements [6,7] shows that humanity and, in particular, the current generation has the power to shape the planet in ways that influence the living conditions for many generations to come. Many renowned scholars think that climate change *mitigation* by a rapid decarbonization of the global social metabolism is the only way to avoid large-scale suffering for many generations, and some suggest a "carbon law" by which global greenhouse gas emissions must be halved every decade from now [8] to achieve United Nations Sustainable Development Goals within planetary boundaries. Others argue that such a profound transformation of our economy would lead to unacceptable suffering at least in some world regions as well, at least temporarily, and suggest that instead of focusing on mitigation, the focus should be on economic development so that continued economic growth will enable future generations to *adapt* to climate change. Still others advocate trying to avert some negative impacts of climate change by large-scale technological interventions aiming at "climate engineering" [9,10].

Since a later voluntary or involuntary phase-out of many climate engineering measures can have even more disruptive effects than natural tipping elements [11], one should, of course, also be concerned that a focus on climate engineering and, maybe to a somewhat lower degree, also a focus on adaptation might increase humanity's dependence on large-scale infrastructure and fragile technology to much higher levels than we learned to deal with, posing a growing risk of not being able to manage these systems forever.

While one might argue that there does not need to be a strict choice between either mitigation or adaptation, the presence of tipping elements in both the natural Earth system and in social systems [12], and the likelihood of nonlinear feedback loops between them [13], suggests that only significant mitigation efforts will avoid natural tipping, and only significant socio-economic measures will cause the "social" tipping into a decarbonized world economy that is no longer fundamentally based on the combustion of fossil fuels. This means the current generation may face a mainly qualitative rather than a quantitative choice: do or do not initiate a rapid decarbonization? Additionally, this choice might take the form of a dilemma where we can either pursue our development and adaptation pathway and put many generations to come at a persistent risk of technological or management failure, or get on a transformation pathway that sacrifices part of the welfare of one or a few generations to enable all later generations to prosper at much lower levels of risk.

While all this might seem exaggerated, we believe that as long as there is a non-negligible possibility that, indeed, we face such a dilemma, it is worthwhile thinking about its implications, in particular its ethical consequences for the current generation. The contribution we aim at making in this article is, hence, not a descriptive one such as trying to assess policy options or other aspects of humanity's agency, as in integrated assessment modeling [14], or their biospherical impacts, as in Earth system modeling, or the dynamics of the Anthropocene that arises from feedbacks between biophysical, socio-metabolic, and socio-cultural processes, as in the emerging discipline of "World-Earth modeling" [2–5,13–15]. Instead, we aim at making a normative contribution that studies some ethical aspects of the described possible dilemma, independently of whether this dilemma really currently exists. To initiate such an ethical debate and allow it to focus on what we think are the most central aspects of the dilemma, we chose to use the method of *thought experiments* (TEs) for this work, a well-established technique in philosophy, in particular in moral philosophy, that studies real-world challenges through the analysis of often extremely simple and radically exaggerated fictitious situations to identify core problems and test ethical principles and theories [16].

In Section 2, we introduce one such TE in two complementary ways, (i) as a formal abstraction of the above-sketched possible dilemma for humanity; and (ii) as a verbal narrative in the style of a parable. We justify the design of the thought experiment further by relating it to (i) a recent classification of the state-space topology of sustainable management of dynamical systems with desirable states [17] and (ii) a very low-dimensional conceptual model of long-term climate and economic development designed to illustrate that classification [18]. In Section 3, we start discussing the ethical aspects of the TE by analyzing it with the tools of rationality-based frameworks, in particular optimal control theory, welfare economics and game theory. This is complemented in Section 4 by a short discussion of alternative approaches based on the notions of responsibility, safe operating spaces, and different sustainability paradigms. Section 5 concludes the paper.

2. A Thought Experiment

Before giving a verbal narrative, we describe our TE in more formal terms, using some simple terminology of dynamical systems theory, control theory and welfare economics:

> *Assume there is a well-defined infinite sequence of generations of humanity, the current one being numbered 0, future ones 1, 2, . . . , and past ones −1, −2, ... At each point in time, one generation is "in charge" and can make choices that influence the "state of the world". The possible states of the world can be classified into just four possible overall states, abbreviated L, T, P, and S, and we assume that this overall state changes only slowly, from generation to generation, due to the inherent*

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

> *dynamics of the world and humanity's choices. We assume the overall state in generation k + 1, denoted X(k + 1), only depends on the following three things: (i) on the immediately preceding state, i.e., that in generation k, denoted X(k); (ii) in some states on the aggregate behavior of generation k, denoted U(k) and called generation k's "choice"; and (iii) in some states also on chance; all this in a way that is the same for each generation (i.e., does not explicitly depend on the generation number k). Being in state X(k) implies a certain overall welfare level for generation k, denoted W(k). We assume the possible choices and their consequences depend on the state X(k) as follows:*
>
> - *Up until generation 0 and including it, all generations have been in state X(k) = L, where welfare is "high", denoted W(k) = 1. When in state X(k) = L, generation k has two choices, A (which is considered the "default" choice that all generations before 0 have made) and B.*
>   - *If generation k chooses option A, the next state is either L or T, depending somewhat on chance. It will be again X(k + 1) = L with some (typically large) probability $\eta > 0$, which is a time-independent constant, and will be X(k + 1) = T with probability $\pi = 1 - \eta > 0$.*
>   - *If they choose option B, the next state will be X(k + 1) = P for sure.*
> - *In state X(k) = T, welfare is low, denoted W(k) = 0, and the state will never change again, X(k′) = T for all k′ > k;*
> - *In state X(k) = P, welfare is also "low", W(k) = 0, but the next state will be X(k + 1) = S for sure; and*
> - *Finally, in state X(k) = S, welfare is again high, W(k) = 1, and the state will never change again, X(k′) = S for all k′ > k.*
>
> *We assume all this is known to generation 0 and all later generations.*

Note that this TE has one free parameter, the probability η. Figure 1 shows this setup. Obviously, one may be immediately tempted to make the TE more "realistic" by introducing additional aspects, such as overlapping generations, a finer distinction between states, options, or welfare levels, more than one "decision-maker", more possible transitions, or even an explicit time dependency to account for external factors. However, we boldly abstain from doing any of that at this point to keep the situation as simple as possible, allowing us to focus only on those aspects present in the TE for our analysis. Rather than justifying what we ignored, we will justify what we put into the TE, but only after having given a verbal, parable-like version of the TE:

> *On an island very far away from any land lives a small tribe whose main food resource are the fruits of a single ancient big tree despite which only grass grows on the island. Although the tree is so strong that it would never die from natural causes, every year there is a rainy season with strong storms, and someday one such storm might kill and blow away the tree. In fact, until just one generation ago, there was a second such tree that was blown away during a storm. If the same happens to the remaining tree, the tribe would have to live on grass forever, having no other food resource. Every generation so far has passed down the knowledge of a rich but unpopulated land across the large sea that can be safely reached if they build a large and strong boat from the tree's trunk. Still, the tribe is so small and the journey would be so hard that they would have to send all their people to be sure the journey succeeds. Also, the passage would take so much time that a whole generation would have to live aboard and hope to catch the odd fish for food, causing deep suffering, and would not be able to see the new land with their own eyes, only knowing their descendants would live there happily and safely for all generations to come. No generation has ever set off on this journey.*

The main purpose of this narrative is not to add detail to the TE, but only to make it more accessible by suggesting a possible alternative interpretation of the states and options in the experiment that is simpler than the actual application to humanity and the Earth system that we motivated it with originally in the introduction. As any such narrative contains details that are not central to the problem one wants to study, but which might distract the analysis, the existence of two alternative narratives may also be used to check which aspects of them are actually crucial elements of the TE (namely those occurring in both narratives) and which are not. While the following text may sometimes refer to either narrative, our analysis will only depend on the formal specification.

Now why did we choose the specific formal specification above? The main justification is that it is essentially the form the potential decision dilemma between adaptation/growth and mitigation sketched in the introduction takes when one uses the recently developed theory of the state-space (rather than geographic) *topology* of sustainable management (TSM, [17]) to analyze a conceptual model of long-term climate and economic development [18].

TSM is a classification of the possible states of a dynamical system (such as the coupled system of natural Earth and humans on it) which has both a default dynamics (which it will display without the interference of an assumed "decision-maker" or "manager" such as a fictitious world government) and a number of alternative, "managed" dynamics (which the decision-maker may bring about by making certain "management" choices). The TSM classification starts with such a system and a set of possible states considered "desirable", and then classifies each possible system state with regard to questions, such as "is this state desirable", "will the state remain desirable by default/by suitable management", "can a desirable state be reached with/without leaving the desirable region", etc. This results in a number of state space regions that differ qualitatively w.r.t. the possibility of sustainable management. One of the most important among these state space regions is what is called a "lake" in TSM. In a "lake", the decision-maker faces the dilemma of either (i) moving the system into an ultimately desirable and secure region called a "shelter", but having to cross an undesirable region to do so; or (ii) using suitable management to avoid ever entering the undesirable region as long as management is sustained, but knowing that the system will enter the undesirable region when management is stopped, which leaves a permanent risk and makes the lake region insecure. Rather than giving the mathematical details of TSM (see [17] for those), let us exemplify these notions with a simple model of long-term climate and economic development, which was analyzed with TSM in [18].

The "AYS" model is a very simple conceptual model of long-term global climate and economic development, describing the deterministic development of just three aggregate continuous variables in continuous time via the ordinary differential equations:

$$dA/dt = E - A/\tau_A$$

$$dY/dt = (\beta - \theta\ A)\ Y$$

$$dS/dt = R - S/\tau_S$$

with the auxiliary quantities:

$$E = F/\varphi,\ F = G\ U,\ R = (1 - G)\ U,\ U = Y/\varepsilon,\ G = 1/(1 + (S/\sigma)^\rho).$$

In this, A is the excess atmospheric carbon stock over preindustrial levels, naturally decaying towards zero at rate τ_A but growing due to emissions E; Y is the gross world economic product, growing at a basic rate β slowed by climate-related damages; θ is the sensitivity of this slowing to A; S is the global knowledge stock for producing renewable energy, decaying at rate τ_S, but growing due to learning-by-doing in proportion to produced renewable energy R; energy efficiency ε stays constant so that total energy use, U, is proportional to Y; energy is supplied by either fossils, F, or renewables, R, in proportions depending on relative price G; σ is the break-even level of S at which fossils and renewables cost the same; ρ is a learning curve exponent; and, finally, emissions are proportional to fossil combustion with combustion efficiency φ.

In [18], several things are shown about this model system: (i) with plausible estimates of the initial state and parameters, it will eventually both violate the climate planetary boundary and stay at welfare levels below current welfare, converging to a fixed point with S = 0; (ii) The system can be forced to neither violate the climate planetary boundary nor to decrease welfare below current levels if humanity has the option to adjust the economic growth rate in real-time within some reasonable levels, but will return to case (i) once this management is stopped; and (iii) if one does not wait too long, it can also be forced to an alternative attractor where S and Y grow indefinitely if humanity can reduce σ by subsidizing renewables or taxing fossils to a reasonable extent, and this management can be phased-out some time after fossils have become uncompetitive, but this decarbonization transition cannot avoid decreasing welfare below current levels for a small number of generations.

In terms of the TSM classification, the attractor where the variables S and Y grow indefinitely lies in a "shelter" region where no management is necessary, and it corresponds to the TE's state "S". The initial state turns out to be in a "lake" region and corresponds to the TE's state "L", while the region one has to cross to reach the shelter from the lake is the state "P" (passage) in the TE. The permanently-managed alternative attractor at which S = 0 corresponds to what TSM calls a "backwater" from which the shelter can no longer be reached. The default attractor with thee planetary boundary and welfare boundary violated is either in what TSM calls a "dark downstream" region since one may still reach the backwater by management, or, if management options have broken down forever, it is in a "trench" region where no escape is possible any longer. If no management is used, the system will move from the lake to the dark downstream which becomes a trench when the management option is removed. In designing our TE, we omitted the dark downstream and simplified the situation so that the system directly goes to the trench ("T") when management breaks down in "L".

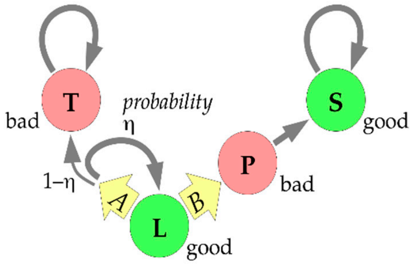


Figure 1. Formal version of the thought experiment. A generation in the good state "L" can choose path "B", surely leading to the good state "S" via the bad state "P" within two generations, or path "A", probably keeping them in "L", but possibly leading to the bad state "T".

3. Analyses Using Rationality-Based Frameworks

We will now start to analyze the ethical aspects of the TE by applying a number of well-established frameworks based on a common assumption of rationality, where we take a broad working definition of rationality here that considers a decision-maker's choice as rational if the decision-maker knows of no alternative choice that gives her a strictly more-preferred prospect than the choice taken, in view of her knowledge, beliefs, and capabilities.

Since we want to focus on what is the ethically right response to the dilemma rather than what makes a politically feasible or implementable choice, we will first treat humanity as a whole as formally just one single infinitely-lived decision-maker that perfectly knows the system as specified in the formal version of the TE, can make a new choice at every generation, can employ randomization for this if desired, can plan ahead, and has the overall goal of having high welfare in all generations. The natural framework for this kind of problem is the language of optimal control theory. Since it will turn out that optimal choices and plans (called "policies" in that language) will very much depend on the evaluation of trajectories (sequences of states) in terms of desirability, we will use concepts such as

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

time preferences, inequality aversion, and risk aversion from decision theory and welfare economics to derive candidate intergenerational welfare functions to be used for this evaluation, and will discuss their impact on the optimal policy. We will restrict our analysis to a consequentialist point of view that takes into account only the actual and potential consequences of actions and their respective probabilities, and leave the inclusion of non-consequentialist, e.g., procedural [19], preferences for later work.

After that, we will refine the analysis by considering each generation a new decision-maker, so that humanity can no longer plan its own future choices, but rather a generation can only recommend and/or anticipate later generations' choices. The natural framework for this kind of problem is the language of game theory. While most of economic theory applies game theory to selfish players, we will apply it instead to players with social preferences based on welfare measures since, in our TE, a generation's welfare is deliberately assumed to be independent of their own choice between A and B.

3.1. Optimal Control Framework with Different Intergenerational Welfare Functions

3.1.1. Terminology

A *trajectory*, X, is a sequence of states X(0), X(1), . . . in the set {L, T, P, S}, where X(t) specifies the state generation t will be in. The only possible trajectories in our TE are

- "XcLT" = (L, . . . , L, T, T, . . .), with c > 0 times L and then T forever (so that c is the time of "collapse");
- "XkLPS" = (L, . . . , L, P, S, S, . . .), with k > 0 times L, then once P, then S forever; and
- "all-L" = (L, L, . . .), which is possible, but has a probability of zero.

A *reward sequence (RS*, sometimes also called a payoff stream), denoted r, is a sequence r(0), r(1), r(2), . . . in the set {0, 1}, where r(t) = 0 or 1 means generation t has low or high overall welfare, respectively. Each trajectory determines an RS via r(t) = 1 if X(t) in {L, S} and r(t) = 0 otherwise. The only possible RSs are, thus:

- "rc10" = (1, . . . , 1, 0, 0, . . .) with c > 0 ones and then zeros forever;
- "rk101" = (1, . . . , 1, 0, 1, 1, . . .) with k > 0 ones, then one zero, then ones forever; and
- "all-1" = (1, 1, . . .), which is possible, but has a probability of zero.

A (randomized) *policy* (sometimes also called a strategy) from time 0 on, denoted as p, is just a sequence of numbers p(0), p(1), p(2), . . . in the interval [0, 1], where p(t) specifies the probability with which generation t will choose option A (staying in L) if they are in state L, i.e., if X(t) = L. In view of the possible trajectories, we may, without loss of generality, assume that if p(t) = 0 for some t, all later entries are irrelevant since state L will never occur after generation t. Thus, we consider only policies of the form:

- infinite sequences (p(0), p(1), . . .) with all p(t) > 0,
- finite sequences (p(0), p(1), . . . , p(k − 1), 0) with p(t) > 0 for all t < k.

The two most extreme ("polar") policies are:

- "all-A" = (1, 1, . . .),
- "directly-B" = (0),

and another interesting set of policies is:

- "Bk" = (1, 1, . . . , 1, 0) with k + 1 ones, where the case k = 0 is "directly-B" and k $\to \infty$ is "all-A",

all of which are deterministic. A policy p is *time-consistent* iff it is a Markov policy, i.e., if and only if all its entries p(t) are equal, so the only time-consistent policies are "all-A", "directly-B", and the policies:

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

- "Ax" = (x, x, x, . . .) with $0 < x < 1$, where the case $x \to 0$ is "directly-B" and $x \to 1$ is "all-A".

Given a policy p, the possible trajectories and RSs have these probabilities:

- $P(XcLT \mid p) = P(rc10 \mid p) = p(0)\,\eta\, p(1)\,\eta \ldots p(c-2)\,\eta\, p(c-1)\,\pi$
- $P(XkLPS \mid p) = P(rk101 \mid p) = p(0)\,\eta\, p(1)\,\eta \ldots p(k-2)\,\eta\,(1 - p(k-1))$
- $P(\text{all-L} \mid p) = P(\text{all-1} \mid p) = 0$

Thus, each policy p defines a probability distribution over RSs, called a *reward sequence lottery (RSL)* here, denoted as RSL(p).

The only missing part of our control problem specification is now a function that numerically evaluates RSLs, or some other information on what RSLs are preferred over which others, in a way that allows the derivation of optimal policies. Let us assume we have specified a *binary social preference relation* that decides for each pair of RSLs g, h which of the following four cases holds: (i) g is strictly better than h, denoted as $g > h$; (ii) the other way around, $h > g$; (iii) they are equally desirable, $g \sim h$; or (iv) they are incomparable, denoted as $g \mid h$. We use the abbreviation $g \geq h$ for $g > h$ or $g \sim h$, and $g \leq h$ for $g < h$ or $g \sim h$. For example, we might put $g > h$ iff $V(g) > V(h)$ and $g \sim h$ iff $V(g) = V(h)$ for some evaluation function V.

Let us assume the social preference relation has the "consistency" property that each non-empty set C of RSLs contains some g such that $h > g$ for no h in C. Then for each non-empty set C of policies, we can call any policy p in C *optimal under the constraint C* (or *C-optimal* for short) iff $RSL(q) > RSL(p)$ for no q in C. In particular, if the preference relation encodes ethical desirability, C contains all policies deemed ethically acceptable, and p is C-optimal, then generation 0 has a good ethical justification in choosing option A with probability p(0) and option B with probability $1 - p(0)$.

We will now discuss several such preference relations and the resulting optimal policies. A common way of assessing preferences over lotteries is by basing them on preferences over certain outcomes, hence, we first consider whether each of two certain RSs, r and s, is preferable. A minimal plausible preference relation is based only on the *Pareto principle* that $r(t) \geq s(t)$ for all t should imply $r \geq s$, and $r(t) \geq s(t)$ for all t but $r(t) > s(t)$ for some t should imply $r > s$. In our case, the only strict preferences would then be between the RSs "all-1", "rk101", "rc10", and "rc′10" for $c > c'$, where we would have "all-1" > "rk101" > "rc10" > "rc′10". However, this does not suffice to make policy decisions, e.g., when we just want to compare policies "directly-B" with "all-A", we need to compare RS "rk101" for $k = 1$ with a lottery over RSs of the form "rc10" for all possible values of c.

One possible criterion for preferring r over s is their degree of "sustainability". The literature contains several criteria by which the sustainability of an RS could be assessed (see [20] for a detailed discussion). The *maximin* criterion (also known as the Rawlsian rule) focuses on the lowest welfare level occurring in an RS, which in all our cases is 0, hence, this criterion does not help in distinguishing options A and B. The *satisfaction of basic needs* criterion [21] asks from what time on welfare stays above some minimal level; if we use 1 as that level, this criterion prefers RS (1, 0, 1, 1, . . .) to all other RS that can occur with positive probability in our TE, hence, it will recommend policy "directly-B", since it makes sure that generation 2 on welfare stays high. The *overtaking* and *long-run average* criteria [21] consider all RSs "rk101" equivalent and strictly more sustainable than all RSs "rc10", hence, they also recommend "directly-B" since that is the only policy avoiding permanently low welfare for sure. Other sustainability criteria are based on the idea of aggregating welfare over time, which we will discuss next.

3.1.2. Aggregation of Welfare over Time

Let us now focus on the simple question whether the RS "rB" = (1, 0, 1, 1, . . .) that results from "directly-B" is preferable to the RS "rc10" = (1, 1, . . . , 1, 0, 0, . . .) with c ones, which may result from "all-A"? This may be answered quite differently. The easy way out is to deem them incomparable since, for some time points t, $rB(t) > rc10(t)$, while for other t, $rc10(t) > rB(t)$, but this does not help. A strong argument is that "rB" should be preferred since it has the larger number of generations with high welfare. Still, at least economists would object that real people's evaluations of future prospects are typically subject to *discounting*, so that a late occurrence of low welfare would be considered less harmful than an early one. A very common approach in welfare economics is, therefore, to base the preference over RSs on some quantitative evaluation v(r), called an *intergenerational welfare function*, which in some way "aggregates" the welfare levels in r and can then also be used as a basis of an evaluation function V(g) of RSLs, which further aggregates the evaluations of all possible RSs in view of their probability. However, let us postpone the consideration of uncertainty for now and stick with the two deterministic RSs "rB" and "rc10".

The most commonly used form of discounting (since it can lead to time-consistent choices) is *exponential discounting*, which would make us evaluate any RS r as:

$$v(r) = r(0) + \delta\, r(1) + \delta^2\, r(2) + \delta^3\, r(3) + \ldots ,$$

using powers of a discount factor $0 \leq \delta < 1$ that encodes humanity's "time preferences". For the above "rB" and "rc10", this gives $v(rB) = (1 - \delta + \delta^2)/(1 - \delta)$ and $v(rc10) = (1 - \delta^c)/(1 - \delta)$. Thus, with exponential discounting, "rB" > "rc10" iff $1 - \delta + \delta^2 > 1 - \delta^c$ or, equivalently, $\delta^{c-1} + \delta > 1$, i.e., the policy "directly-B" is preferable iff δ is large enough or c is small enough. Since $1/\delta$ can be interpreted as a kind of (fuzzy) evaluation time horizon, this means that "directly-B" will be preferable iff the time horizon is large enough to "see" the expected ultimate transition to state T at time c under the alternative extreme policy "all-A". At what δ exactly the switch occurs depends on how we take into account the uncertainty about the collapse time c, i.e., how we get from preferences over RSs to preferences over RSLs, which will be discussed later. A variant of the above evaluation v due to Chichilnisky [21] adds to v(r) some multiple of the long-term limit, $\lim_{t\to\infty} r(t)$, which is 1 for "rB" and 0 for all "rc10", thus making "directly-B" preferable also for smaller δ, depending on the weight given to this limit.

Let us shortly consider the alternative policy "Bk" = (1, . . . , 1, 0) with k ones, where choosing B is delayed by k periods, and "B1" equals "directly-B". If k < c, this results in RS r(k + 1)101, which is evaluated as $(1 - \delta^{k+1} + \delta^{k+2})/(1 - \delta)$, which grows strictly with growing k. Thus, if the collapse time c was known, the best policy among the "Bk" would be the one with k = c − 1, i.e., initiating the transition at the last possible moment right before the collapse, which is evaluated as $(1 - \delta^c + \delta^{c+1})/(1 - \delta) > (1 - \delta^c)/(1 - \delta)$, hence, it would be preferred to "all-A". However, c is, of course, not known, but a random variable, so we need to come back to this question when discussing uncertainty below.

An argument against exponential discounting is that even for values of δ close to 1, late generations' welfare would be considered too unimportant. Under the most common alternative form of discounting, *hyperbolic discounting*, one would instead have the evaluation:

$$v(r) = r(0) + r(1)/(1 + \kappa) + r(2)/(1 + 2\kappa) + r(3)/(1 + 3\kappa) + \ldots$$

with some positive constant κ. Hyperbolic discounting can easily be motivated by an intrinsic suspicion that, due to factors unaccounted for, the expected late rewards may not actually be realized, but that the probability of this happening is unknown and has to be modeled via a certain prior distribution [22]. Under hyperbolic discounting, v(rB) is infinite while v(rc10) is finite independently of k, so the policy "directly-B" would always be preferable to "all-A" no matter how uncertainty about the actual c is accounted for.

A somewhat opposite alternative to hyperbolic discounting is what one could call "rectangular" discounting: simply average the welfare of only a finite number, say H many, of the generations:

$$v(r) = (r(0) + \ldots + r(H - 1))/H,$$

where H is the evaluation horizon. With this, $v(rB) = (H - 1)/H$ and $v(rc10) = \min(c, H)/H$, so that $v(rB) > v(rc10)$ iff $H > c + 1$. Thus, again, "directly-B" is preferable if the horizon is large enough.

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

3.1.3. Social Preferences over Uncertain Prospects: Expected Probability of Regret

Let us now consider evaluations of RSLs rather than RSs, which requires us to take into account the probabilities of all possible RSs that an RSL specifies.

If we already have a social preference relation "$\geq$" on RSs, such as one of those discussed above, then a very simple idea is to consider an RSL g'' strictly preferable to another RSL g' iff the probability that a realization $r''(g'')$ of the random process g'' is strictly preferable to an independent realization $r'(g')$ of the random process g' is strictly larger than $1/2$:

$$g'' > g' \text{ iff } P(r''(g'') > r'(g')) > 1/2.$$

The rationale for this is based in the idea of *expected probability of regret*. Assume policy p was chosen, resulting in some realization r(RSL(p)), and someone asks whether not policy q should have been taken instead and argues that this should be evaluated by asking how likely the realization r′(RSL(q)) under the alternative policy would have been strictly preferable to the actual realization r(RSL(p)). Then the probability of the latter, averaged over all possible realizations r(RSL(p)) of the policy actually taken, should be not too large. This expected probability of regret is just $P(r''(g'') > r'(g'))$ for $g' = \text{RSL}(p)$ and $g'' = \text{RSL}(q)$. Since for the special case where $g' = g''$, the value $P(r''(g'') > r'(g'))$ can be everything up to at most $1/2$, the best we can hope for is that $P(r''(\text{RSL}(q)) > r'(\text{RSL}(p))) \leq 1/2$ for all $q \neq p$ if we want to call p optimal.

In our example, the polar policy "directly-B" results in an RSL "gB" which gives 100% probability to RS "rB", the opposite polar policy "all-A" results in an RSL "gA" which gives a probability of $\eta^{c-1}\pi$ to RS "rc10", and other policies result in RSLs with more complicated probability distributions. e.g., with exponential discounting, rB > rc10 iff $\delta^{c-1} + \delta > 1$, hence, "gB" > "gA" iff the sum of $\eta^{c-1}\pi$ over all c with $\delta^{c-1} + \delta > 1$ is larger than $1/2$. If $c(\delta)$ is the largest such c, which can be any value between 1 (for $\delta \to 0$) and infinity (for $\delta \to 1$), that sum is $1 - \eta^{c(\delta)}$, which can be any value between π (for $\delta \to 0$) and 1 (for $\delta \to 1$). Similarly, with rectangular discounting, "rB" > "rc10" iff $H > c + 1$, hence, "gB" > "gA" iff $1 - \eta^{H-1} > 1/2$. In both cases, if $\eta < 1/2$, "directly-B" is preferred to "all-A", while for $\eta > 1/2$, it depends on δ or H, respectively. In contrast, under hyperbolic discounting, "directly-B" is always preferred to "all-A".

What about the alternative policy "Bk" as compared to "all-A"? If $c \leq k$, we get the same reward sequence as in "all-A", evaluated as $(1 - \delta^c)/(1 - \delta)$. If $c > k$, we get an evaluation of $(1 - \delta^{k+1} + \delta^{k+2})/(1 - \delta)$, which is larger than $(1 - \delta^c)/(1 - \delta)$ iff $\delta^{c-k-1} + \delta > 1$. Thus, RSL(Bk) > gA iff the sum of $\eta^{c-1}\pi$ over all $c > k$ with $\delta^{c-k-1} + \delta > 1$ is larger than $1/2$. Since the largest such c is $c(\delta) + k$, that sum is $\eta^k(1 - \eta^{c(\delta)})$, so whenever "Bk" is preferred to "all-A", then so is "directly-B". Let us also compare "Bk" to "directly-B". In all cases, "directly-B" gets $(1 - \delta + \delta^2)/(1 - \delta)$, while "Bk" gets the larger $(1 - \delta^{k+1} + \delta^{k+2})/(1 - \delta)$ if $c > k$, but only $(1 - \delta^c)/(1 - \delta)$ if $c \leq k$. The latter is $< (1 - \delta + \delta^2)/(1 - \delta)$ iff $c \leq c(\delta)$. Thus, "directly-B" is strictly preferred to "Bk" iff $1 - \eta^{\min(c(\delta),k)} > 1/2$, i.e., iff both $c(\delta)$ and k are larger than $\log(1/2)/\log(\eta)$, which is at least fulfilled when $\eta < 1/2$. Conversely, "Bk" is strictly preferred to "directly-B" iff either $c(\delta)$ or k is smaller than $\log(1/2)/\log(\eta)$. In particular, if social preferences were based on the expected probability of regret, delaying the choice for B by at least one generation would be strictly preferred to choosing B directly whenever $\eta > 1/2$, while at the same time, delaying it forever would be considered strictly worse at least if the time horizon is long enough. Basing decisions on this maxim would, thus, lead to time-inconsistent choices: in every generation, it would seem optimal to delay the choice B by the same positive number of generations, but not forever, so no generation would actually make that choice.

Before considering a less problematic way of accounting for uncertainty, let us shortly discuss a way of deriving preferences over RSs rather than RSLs that is formally similar to the above. In that case the rationale would not be in terms of regret but in terms of Rawls' *veil of ignorance*. Given two

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

RSs r' and r'', would one rather want to be born into a randomly selected generation in situation r' or into a randomly selected generation in situation r''? i.e., let us put:

$$r'' > r' \text{ iff } P(r''(t'') > r'(t')) > 1/2,$$

where t'', t' are drawn independently from the same distribution, e.g., the uniform one on the first H generations or a geometric one with parameter δ. Then $rB(t'') > rc10(t')$ iff $rB(t'') = 1$ and $rc10(t') = 0$, i.e., iff $t'' \neq 1$ and $t' > c$. Under the uniform distribution over H generations, the latter has a probability of $(H - 1)(H - c)/H^2$ if $H \geq c$, which can be any value between 0 (for $H = c$) and 1 (for very large H), hence, whether "rB" > "rc10" depends on H again. Similarly, $rB(t'') < rc10(t')$ iff $rB(t'') = 0$ and $rc10(t') = 1$. This has probability $\min(c,H)/H^2$, which is 1 for $H = 1$ and approaches 0 for very large H, hence, whether "rB" < "rc10" depends on H as well. However, this version of preferences over RS leaves a large possibility for undecidedness, "rB" | "rc10", where neither "rB" > "rc10" nor "rc10" > "rB". This is the case when both $(H - 1)(H - c)/H^2$ and $\min(c,H)/H^2$ are at most 1/2, i.e., when $\max[(H - 1)(H - c), \min(c,H)] \leq H^2/2$, which is the case when $H \geq 2$ and $H^2 - 2(c + 1)H + 2c \leq 0$, i.e., when $2 \leq H \leq c + 1 + (c^2 + 1)^{1/2}$. A similar result holds for the geometric distribution with parameter δ. Thus, while the probability of regret idea can lead to time-inconsistent choices, the formally similar veil of ignorance idea may not be able to differentiate enough between choices. Another problematic property of our veil of ignorance-based preferences is that they can lead to preference cycles. e.g., assume $H = 3$ and compare the RSs $r = (0, 1, 2)$, $r' = (2, 0, 1)$, and $r'' = (1, 2, 0)$. Then it would occur that $r > r' > r'' > r$, so there would be no optimal choice among the three.

3.1.4. Evaluation of Uncertain Prospects: Prospect Theory and Expected Utility Theory

We saw that the above preference relations based on regret and the veil of ignorance, while intuitively appealing, are, however, unsatisfactory from a theoretical point of view, since they can lead to time-inconsistent choices and preference cycles, i.e., they may fail to produce clear assessments of optimality. The far more common way of dealing with uncertainty is, therefore, based on numerical evaluations instead of binary preferences. A general idea, motivated by a similar theory regarding individual, rather than social, preferences, called *prospect theory* [23], is to evaluate an RSL g by a linear combination of some function of the evaluations of all possible RSs r with coefficients that depend on their probabilities:

$$V(g) = \textstyle\sum_r w(P(r \mid g))\, f(v(r)).$$

In the simplest version, corresponding to the special case of *expected utility theory*, both the probability weighting function w and the evaluation transformation function f are simply the identity, $w(p) = p$ and $f(v) = v$, so that $V(g) = \sum_r P(r \mid g)\, v(r) = E_g\, v(r)$, the expected evaluation of the RSs resulting from RSL g. If combined with a $v(r)$ based on exponential discounting, this gives the following evaluations of our polar policies:

$$V(\text{RSL(directly-B)}) = v(rB) = (1 - \delta + \delta^2)/(1 - \delta)$$

and:

$$V(\text{RSL(all-A)}) = E_{\text{all-A}}\, v(rc10) = \textstyle\sum_{c>0} \eta^{c-1}\pi(1 - \delta^c)/(1 - \delta) = 1/(1 - \delta\eta).$$

Hence, "directly-B" is preferred to "all-A" iff $(1 - \delta + \delta^2)(1 - \eta\delta) - 1 + \delta > 0$. Again, this is the case for $\delta > \delta_{\text{crit}}(\eta)$ with $\delta_{\text{crit}}(0) = 0$ and $\delta_{\text{crit}}(1) = 1$. The result for rectangular discounting is similar, while for hyperbolic discounting "directly-B" is always preferred to "all-A", and all of this as expected from the considerations above.

In prospect theory, the transformation function f can be used to encode certain forms of risk attitudes. For example, we could incorporate a certain form of risk aversion against uncertain social welfare sequences by using a strictly concave function f, such as $f(v) = v^{1-a}$ with $0 < a < 1$ (isoelastic case) or $f(v) = -\exp(-av)$ with $a > 0$ (constant absolute risk aversion) (welfare economists might be confused a little by our discussion of risk aversion since they are typically applying the concept in the context of consumption, income or wealth of individuals at certain points in time, in which context one can account for risk aversion already in the specification of individual consumers' utility function, e.g., by making utility a concave function of individual consumption, income, or wealth. Here we are, however, interested in a different aspect of risk aversion, where we want to compare uncertain streams of societal *welfare* rather than uncertain consumption bundles of individuals. Thus, even if our assessment of the welfare of each specific generation in each specific realization of the uncertainty about the collapse time c already accounts for risk aversion in individual consumers in that generation, we still need to incorporate the possible additional risk aversion in the "ethical social planner"). This basically leads to a preference for small variance in v. One can see numerically that in both cases increasing the degree of risk aversion, a, lowers $\delta_{\text{crit}}(\eta)$, not significantly so in the isoelastic case but significantly in the constant absolute risk aversion case, hence, risk aversion favors "directly-B". In particular, the constant absolute risk aversion case with $a \to \infty$ is equivalent to a "worst-case" analysis that always favors "directly-B". Conversely, one can encode risk-seeking by using $f(v) = v^{1+a}$ with $a > 0$.

Under expected utility theory, the delayed policy "Bk" has:

$$(1-\delta) \times V(RSL(Bk)) = \eta^k(1-\delta^{k+1}+\delta^{k+2}) + \sum\nolimits_{c=1\ldots k} \eta^{c-1}\pi(1-\delta^c),$$

which is either strictly decreasing or strictly increasing in k. Since "directly-B" and "all-A" corresponds to the limits $k \to 0$ and $k \to \infty$, "Bk" is never optimal but always worse than either "directly-B" or "all-A". The same holds with risk-averse specifications of f. Under isoelastic risk-*seeking* with $f(v) = v^{1+a}$, however, we have:

$$(1-\delta) \times V(RSL(Bk)) = \eta^k(1-\delta^{k+1}+\delta^{k+2})^{1+a} + \sum\nolimits_{c=1\ldots k} \eta^{c-1}\pi(1-\delta^c)^{1+a},$$

which may have a global maximum for a strictly positive and finite value of k, so that delaying may seem preferable. e.g., with $\delta = 0.8$, $\eta = 0.95$, and $a = 1/2$, V(RSL(Bk)) is maximal for $k = 6$, i.e., one would want to choose six times A before choosing B, again a time-inconsistent recommendation.

As long as the probability weighting function w is simply the identity, there is always a deterministic optimal policy. While other choices for w could potentially lead to non-deterministic optimal policies, they can be used to encode certain forms of risk attitudes that cannot be encoded via f. e.g., one can introduce some degree of optimism or pessimism by over- or underweighing the probability of the unlikely cases where c is large. For example, if we put $w(p) = p^{1-b}$ with $0 \leq b < 1$, then increasing the degree of optimism b, one can move $\delta_{\text{crit}}(\eta)$ arbitrarily close towards 1, which is not surprising. We will however not discuss this form of probability reweighting further but will use a different way of representing "caution" below. Since that form is motivated by its formal similarity to a certain form of inequality aversion, we will discuss the latter first now before returning to risk attitudes.

3.1.5. Inequality Aversion: A Gini-Sen Intergenerational Welfare Function

While discounting treats different generations' welfare differently, it only does so based on time lags, and all the above evaluations still only depend on some form of (weighted) time-average welfare and are blind to welfare *inequality* as long as these time-averages are the same. However, one may argue that an RS with less inequality between generations, such as (1, 1, 1, . . .), should be strictly preferable to one with the same average but more inequality, such as (2, 0, 2, 0, 2, 0, . . .). Welfare economics has come up with a number of different ways to make welfare functions sensitive to inequality, and although most of them were initially developed to deal with inequality between individuals of a society at a given point in time (which we might call "intragenerational" inequality here), we can use the same ideas to deal with inequality between welfare levels of different generations ("intergenerational" inequality). Since our basic welfare measure is not quantitative but qualitative since it only distinguishes "low" from "high" welfare, inequality metrics based on numerical transformations, such as the Atkinson-Theil-Foster family of indices, are not applicable in our context, but the Gini-Sen welfare function [24], which only requires an ordinal welfare scale, is. The idea is that the value of a specific allocation of welfare to all generations is the expected value of the smaller of the two welfare values of two randomly-drawn generations. If the time horizon is finite, H > 0, this leads to the following evaluation of an RS r:

$$V_2(r) = (\sum_{t=0\ldots H-1} \sum_{t'=0\ldots H-1} \min[r(t), r(t')])/H^2.$$

It is straightforward to generalize the idea from drawing two to drawing any integer number $a > 0$ of generations, leading to a sequence of welfare measures $V_a(r)$ that get more and more inequality averse as a is increased from 1 (no inequality aversion, "utilitarian" case) to infinity (complete inequality aversion), where the limit for $a \to \infty$ is the *egalitarian* welfare function:

$$V_1(r) = [r(0) + \ldots + r(H-1)]/H$$

$$V_a(r) = (\sum_{t1=0\ldots H-1} \cdots \sum_{ta=0\ldots H-1} \min[r(t_1), \ldots, r(t_a)])/H^a$$

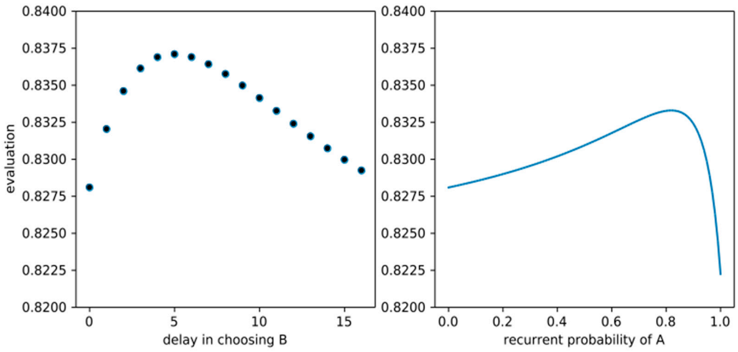
$$V_\infty(r) = \min[r(0), \ldots, r(H-1)]$$

Note that $I = 1 - V_2(r)/V_1(r)$ is the Gini index of inequality and the formula $V_2(r) = V_1(r)\,(1 - I)$ is often used as the definition of the Gini-Sen welfare function.

Our RSs "rc10" then gets $V_a(\text{rc10}) = \min(c/H, 1)^a$, while "rk101" gets $V_a(\text{rk101}) = [(H-1)/H]^a$ if $k < H$ and $V_a(\text{rk101}) = 1$ if $k \geq H$. Together with expected utility theory for evaluating the risk about c, this makes:

$$V_a(\text{all-A}) = \eta^H + \sum_{c=1\ldots H} \eta^{c-1}\pi(c/H)^a$$

and $V_a(\text{directly-B}) = [(H-1)/H]^a$. Numerical evaluation shows that even for large H, "all-A" may still be preferred due to the possibility that collapse will not happen before H and all generations will have the same welfare, but this is only the case for extremely large values of a. If we use exponential instead of rectangular discounting and compare the policies "directly-B", "Bk", and "all-A", we may again get a time-inconsistent recommendation to choose B after a finite number of generations. e.g., Figure 2a shows V(Bk) vs. k for the case $\eta = 0.985$, $\delta = 0.9$, $a = 2$, where the optimal delay would appear to be five generations. If we restrict our optimization to the time-consistent policies "Ax", the optimal x in that case would be ≈0.83, i.e., each generation would choose A with about 83% probability and B with about 17% probability, as shown in Figure 2b. Still, note that the absolute evaluations vary only slightly in this example.



(**a**) (**b**)

Figure 2. Inequality-averse evaluation of deterministic but delayed policies (**a**) and time-consistent but probabilistic policies (**b**) for the case $\eta = 0.985$, $\delta = 0.9$ and $a = 2$.

Let us see what effect a formally similar idea has in the context of risk aversion.

3.1.6. Caution: Gini-Sen Applied to Alternative Realizations

What happens if instead of drawing $a \geq 1$ many generations $t_1, \ldots, t_a$ at random, we draw $a \geq 1$ many realizations $r_1, \ldots, r_a$ of an RSL g at random and use the expected minimum of all the RS-evaluations $V(r_i)$ as a "cautious" evaluation of the RSL g?

$$V_a(g) = \sum_{r1} \ldots \sum_{ra} g(r_1) \times \ldots \times g(r_a) \times \min[v(r_1), \ldots, v(r_a)].$$

For a = 1, this is just the expected utility evaluation of g, while for $a \to \infty$, it gives a "worst-case" evaluation. For actual numerical evaluation, the following equivalent formula is more useful (assuming that all $v(r) \geq 0$):

$$V_a(g) = \int_{x \geq 0} P_g(v(r) \geq x)^a \, dx,$$

where $P_g(v(r) \geq x)$ is the probability that $v(r) \geq x$ if r is a realization of g. In that form, a can be any real number ≥ 1 and it turns out that the evaluation is a special case of *cumulative* prospect theory [23], with the cumulative probability weighting function $w(p) = p^a$. Focusing on "all-A" vs. "directly-B" again, we get $V_a(\text{all-A}) = (1 - \eta^{aH})/(1 - \eta^a)H$ and $V_a(\text{directly-B}) = (H - 1)/H$, hence, "all-A" is preferred iff $(1 - \eta^{aH})/(H - 1) > 1 - \eta^a$, i.e., iff H and a are small enough and η is small enough. In particular, regardless of H and η, for $a \to \infty$ we always get a preference for "directly-B" as in the constant absolute risk aversion. This is because with the Gini-Sen-inspired specification of caution, the degree of risk aversion effectively acts as an exponent to the survival probability η, i.e., increasing risk aversion has the same effect as increasing collapse probability, which is an intuitively appealing property.

3.1.7. Fairness as Inequality Aversion on Uncertain Prospects

Consider the RSs r1 = (1, 0, 1) and r2 = (1, 1, 0), and the RSL g that results in r1 or r2 with equal probability 1/2. If we apply inequality aversion on the RS level as above, say with a = 2, we get V(r1) = V(r2) = V(g) = 4/9. Still, g can be considered more *fair* than both r1 and r2 since under g, the expected rewards are (1, 1/2, 1/2) rather than (1, 0, 1) or (1, 1, 0), so no generation is doomed to zero reward but all have a fair chance of getting a positive reward. It is, therefore, natural to consider applying "inequality aversion" on the RSL level to encode fairness, by putting:

$$V_a(g) = \left(\sum_{t1=0 \ldots H-1} \cdots \sum_{ta=0 \ldots H-1} \min[V(g, t_1), \ldots, V(g, t_a)]\right)/H^a,$$

where V(g, t) is some evaluation of the uncertain reward of generation t resulting from g, e.g., the expected reward or some form of risk-averse evaluation. The interpretation is that $V_a(g)$ is the expected minimum of how two randomly drawn generations within the time horizon evaluate their uncertain rewards under g. Using exponential discounting instead, the formula becomes:

$$V_a(g) = (1 - \delta)^a \sum_{t1=0 \ldots H-1} \cdots \sum_{ta=0 \ldots H-1} \delta^{t1+\ldots+ta} \min[V(g, t_1), \ldots, V(g, t_a)].$$

If we use the expected reward for V(g, t) and evaluate the time-consistent policies "Ax" with this $V_a(g)$, the result looks similar to Figure 2b, i.e., the optimal time-consistent policy is again non-deterministic. A full optimization of $V_a(g)$ over the space of all possible probabilistic policies shows that the overall optimal policy regarding $V_a(g)$ is not much different from the time-consistent one, it prescribes choosing A with probabilities between 79% and 100% in different generations for the setting of Figure 2.

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

3.1.8. Combining Inequality and Risk Aversion with Fairness

How could one consistently combine all the discussed aspects into one welfare function? Since a Gini-Sen-like technique of using minima can be used for each of them, it seems natural to base a combined welfare function on that technique as well. Let us assume we want to evaluate the four simple RSLs $g^1, \ldots, g^4$ listed in Table 1 in a way that makes $V(g^1) > V(g^2)$ because the latter is more risky, $V(g^2) > V(g^3)$ because the latter has more inequality, and $V(g^3) > V(g^4)$ because the latter is less fair. Then we can achieve this by applying the Gini-Sen technique several times to define welfare functions $V^0 \ldots V^6$ that represent more and more of our aspects as follows:

- Simple averaging: $V^0(g) = E_r\, E_t\, r(t)$ where $E_r\, f(r)$ is the expectation of $f(r)$ w.r.t. the lottery g and $E_t\, f(t)$ is the expectation of $f(t)$ w.r.t. some chosen discounting weights;
- Gini-Sen welfare of degree a = 3: $V^1(g) = E_r\, E_{t1}\, E_{t2}\, E_{t3} \min\{r(t_1), r(t_2), r(t_3)\}$;
- Overall risk-averse welfare: $V^2(g) = E_{r1}\, E_{r2} \min\{E_t\, r_1(t), E_t\, r_2(t)\}$;
- Fairness-seeking welfare of degree a = 3: $V^3(g) = E_{t1}\, E_{t2}\, E_{t3} \min\{E_r\, r(t_1), E_r\, r(t_2), E_r\, r(t_3)\}$;
- Inequality- and overall risk-averse welf.: $V^4(g) = E_{r1}\, E_{r2} \min\{v^4(r_1), v^4(r_2)\}$ with $v^4(r) = E_{t1}\, E_{t2}\, E_{t3} \min\{r(t_1), r(t_2), r(t_3)\}$;
- Inequality and overall risk index: $I^4(g) = 1 - V^4(g)/V^0(g)$;
- Generational risk averse and fair welfare: $V^5(g) = E_{t1}\, E_{t2}\, E_{t3} \min\{V^5(g, t_1), V^5(g, t_2), V^5(g, t_3)\}$ with $V^5(g, t) = E_{r1}\, E_{r2} \min\{r_1(t), r_2(t)\}$;
- Generational risk and fairness index: $I^5(g) = 1 - V^5(g)/V^0(g)$; and
- All effects combined: $V^6(g) = V^4(g)V^5(g)/V^0(g) = V^0(g)[1 - I^4(g)][1 - I^5(g)]$

The resulting evaluations for $g^1 \ldots g^4$ can be seen in Table 1. We chose a higher degree of inequality-aversion (a = 3) than the degree of risk-aversion (a = 2) so that $V^6(g^2) > V^6(g^3)$ as desired. Applied to our thought experiment, V^6 can result in properly probabilistic and time-inconsistent policy recommendations, as shown in Figure 3 for two example choices of η and discounting schemes. An alternative way of combining inequality and risk aversion into one welfare function would be to use the concept of *recursive* utility [25], which is, however, beyond the scope of this article.

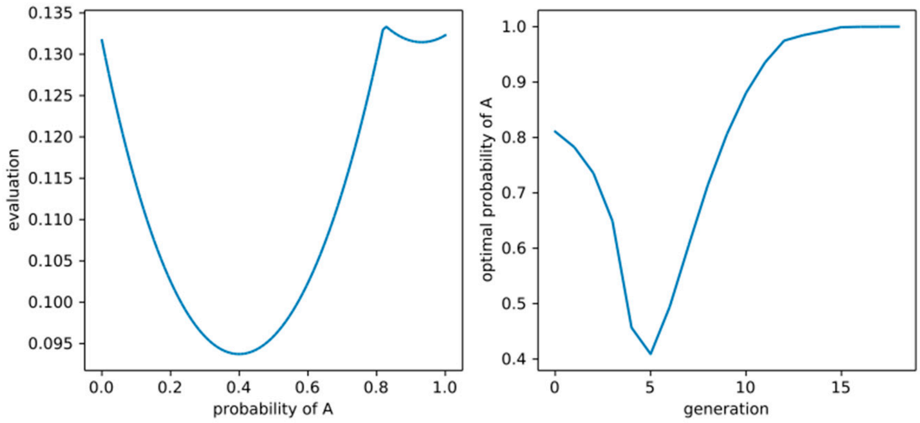


Figure 3. (**Left**) Evaluation V^6 for the case of η = 0.68, rectangular discounting with very short horizon 3 and choosing A for sure in generation 1, by probability of chosing A in generation 0, showing an optimal probability of approximately 82%. (**Right**) Optimal policy for the first 20 generations according to V^6 for the case of η = 0.97 and exponential discounting with δ = 0.9.

Table 1. Comparison of the effects of inequality aversion, overall and generational risk aversion, and fairness on the evaluation of four simple reward sequence lotteries (RSLs). All effects are implemented in the Gini-Sen style (see main text for details), inequality aversion with a larger degree of a = 3, risk aversion and fairness with a lower degree of a = 2, which is reflected in the preference for the coin toss between the "no-inequality" reward sequences (0, 0) and (1, 1) over the coin toss between the "equal average" reward sequences (0, 1) and (1, 0).

RSL	V^0: No Effects	V^1: Only Inequality Aversion	V^2: Only Overall Risk Aversion	V^3: Only Fair-ness	V^4: Inequality and Overall Risk Aversion	V^5: Generational Risk Aversion and Fairness	V^6: All Effects Combined
g^1: (0.5, 0.5) for sure	0.5	0.5	0.5	0.5	0.5	0.5	0.5
g^2: coin toss between (0, 0) and (1, 1)	0.5	0.5	0.25	0.5	0.25	0.25	0.125
g^3: coin toss between (0, 1) and (1, 0)	0.5	0.25	0.5	0.5	0.125	0.25	0.0625
g^4: (0, 1) for sure	0.5	0.25	0.5	0.25	0.125	0.125	0.03125

Summarizing the results of our analysis in the optimal control framework that treats humanity as a single infinitely-lived decision-maker, we see that there is no clear recommendation to either choose A or B at time 0 since depending on the degrees and forms of time preferences/time horizon and risk/inequality/fairness attitudes, either one of the policies "all-A" or "directly-B" may appear optimal, or it may even appear optimal to deterministically delay the choice for B by a fixed number of generations or choose A by a time-varying probability, leading to time-inconsistent recommendations. At least we were able to formally confirm quite robustly the overall intuition that risk aversion and long time horizons are arguments in favor of B while risk seeking and short time horizons are arguments in favor of A. Only the effect of inequality aversion might be surprising, since it can lead to either recommending a time-inconsistent policy of delay (if we restrict ourselves to deterministic policies) or a probabilistic policy of choosing A or B with some probabilities (if we restrict ourselves to time-consistent policies). In the next subsection, we will see what difference it makes that no generation can be sure about the choices of future generations.

3.2. Game-Theoretical Framework

While the above analysis took the perspective of humanity as a single, infinitely lived "agent" that can plan ahead its long-term behavior, we now take the viewpoint of the single generations who care about intergenerational welfare, but cannot prescribe policies for future generations and have to treat them as separate "players" with potentially different preferences instead. For the analysis, we will employ game-theory as the standard tool for such multi-agent decision problems. Each generation, t, is treated as a player who, if they find themselves in state L, has to choose a potentially randomized strategy, p(t), which is, as before, the probability that they choose option A. Since each generation is still assumed to care about future welfare, the optimal choice of p(t) depends on what generation t believes future generations will do if in L. As usual in game theory, we encode these beliefs by subjective probabilities, denoting by q(t′, t) the believed probability by generation t′ that generation t > t′ will choose A when still in L.

Let us abbreviate generation t′ by $G_{t'}$ and the set of generations t > t′ by $G_{>t'}$ and focus on generation t′ = 0 at first. Let us assume that V = V^4, V^5, or V^6 with exponential discounting encodes their social preferences over RSLs. Given $G_{0'}$s beliefs about $G_{>0'}$s behavior, q(0, t) for all t > 0, we then need to find that x in [0, 1] which maximizes V(RSL($p_{x,q}$)), where $p_{x,q}$ is the resulting policy $p_{x,q}$ = (x, q(0, 1), q(0, 2), …). If G_0 believes G_1 will choose B for sure (i.e., q = (0, …) = "directly-B") and chooses strategy x, the resulting RSL($p_{x,q}$) produces the reward sequence r_1 = (1, 0, 0, …) with

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

probability $x\pi$, $r_2 = (1, 0, 1, 1, \dots)$ with probability $1 - x$, and $r_3 = (1, 1, 0, 1, 1, \dots)$ with probability $x\eta$. Hence:

$$\begin{aligned}&V^4(RSL(p_{x,q})) =\\ &\quad x^2\,[(1-(1-\delta)\delta^2)^3\,\eta^2 - (1-\delta)^3\eta^2 + 2(1-\delta)^3\eta - 2(1-(1-\delta)\delta)^3\eta - (1-\delta) + (1-(1-\delta)\delta)^3]\\ &\quad + 2x\,(-(1-\delta)^3\eta + (1-(1-\delta)\delta)^3\eta + (1-\delta)^3 - (1-(1-\delta)\delta)^3] + (1-(1-\delta)\delta)^3.\end{aligned}$$

Since the coefficient in front of x^2 is positive, V^4 is maximal for either $x = 0$, where it is $(1 - \delta + \delta^2)^3$, or for $x = 1$, where it is $(\delta^3 - \delta^2 + 1)^3\eta^2 + (\delta - 1)^3(\eta^2 - 1)$, which is always smaller, so w.r.t. V^4, $x = 0$ (choosing B for sure) is optimal under the above beliefs. For V^5, we have $V^5(RSL(p_{x,q}), t) = 1$ for $t = 0$, $(x\eta)^2$ for $t = 1$, $(1 - x)^2$ for $t = 2$, and $(1 - x\pi)^2$ for $t > 2$. If $x < 1/(1 + \eta)$, we have $(x\eta)^2 < (1 - x)^2 < (1 - x\pi)^2 < 1$, while for $x > 1/(1 + \eta)$, we have $(1 - x)^2 < (x\eta)^2 < (1 - x\pi)^2 < 1$. For $x \leq 1/(1 + \eta)$, $V^5(RSL(p_{x,q}))$ is again quadratic in x with a positive x^2 coefficient with value $1 + (1 - \delta)^3 - (1 - \delta^2)^3$ at $x = 0$ and, again, a smaller value at $x = 1/(1 + \eta)$. Additionally, for $x \geq 1/(1 + \eta)$, $V^5(RSL(p_{x,q}))$ is quadratic in x with positive x^2 coefficient and a value of:

$$1 - \delta(3 - 3\delta + \delta^2 - \eta^2[1 - \delta + \delta^2][3 - \delta(1 - \delta + \delta^2)(3 - \delta + \delta^2 - \delta^3)])$$

for $x = 1$, which is larger than the value for $x = 0$ if η is large enough and/or δ small enough. A similar thing holds for the combined welfare measure V^6, as shown in Figure 4, blue line, for the case $\eta = 0.95$ and $\delta = 0.805$, where G_0 will choose A if they believe G_1 will choose B, resulting in an evaluation $V^6 \approx 0.43$. The orange line in the same plot shows $V^6(RSL(p_{x,q}))$ for the case in which G_0 believes that G_1 will choose A and G_2 will choose B if they are still in L, which corresponds to the beliefs $q = (1, 0, \dots)$. Interestingly, in that case, it is optimal for G_0 to choose A, resulting in an evaluation $V^6 \approx 0.42$. Since the dynamics and rewards do not explicitly depend on time, the same logic applies to all later generations, i.e., for that setting of η and δ and any $t \geq 0$, it is optimal for G_t to choose A when they believe G_{t+1} will choose B and optimal to choose B when they believe G_{t+1} will choose A and G_{t+2} will choose B.

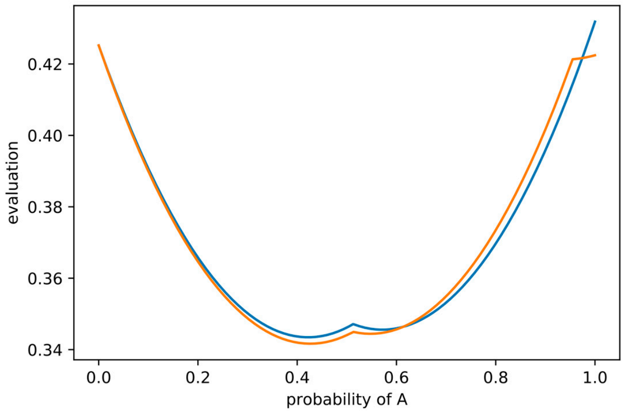


Figure 4. Evaluation V^6 for the case of $\eta = 0.95$ and $\delta = 0.805$ depending on the first generation's probability of choosing A (horizontal axis), for the case where they expect the next generations to choose B (blue line) or to choose A and then B (orange line).

Now assume that all generations have preferences encoded by welfare function V^6 and believe that all generations G_t with even t will choose A and all generations G_t with odd t will choose B. Then it is optimal for all generations to do just that. In other words, these assumed common beliefs form a *strategic equilibrium* (more precisely, a subgame-perfect Nash equilibrium) for that setting of η and δ. However, under the very same set of parameters and preferences, the alternative common belief that all even generations will choose B and all odd ones A also forms such an equilibrium. Another equilibrium consists of believing that all generations choose A with probability ≈83.7% which all generations evaluate as only $V^6 \approx 0.40$, which is less than in the other two equilibria. The existence of more than one strategic equilibrium is usually taken as an indication that the actual behavior is very difficult to predict even when assuming complete rationality. In our case, this means G_0 cannot plausibly defend any particular belief about $G_{>0}'$s policy on the grounds of $G_{>0'}$s rationality since $G_{>0}$ might follow at least either of the three identified equilibria (or still others). In other words, for many values of η and δ a game-theoretic analysis based on subgame-perfect Nash equilibrium might not help G_0 in deciding between A and B. A common way around this is to consider "stronger" forms of equilibrium to reduce the number of plausible beliefs, but this complex approach is beyond the scope of this article. An alternative and actually older approach [26] is to use a different basic equilibrium concept than Nash equilibrium, not assuming players have beliefs about other players policies encoded as subjective probabilities, but rather assuming players apply a worst-case analysis. In that analysis, each player would maximize the minimum evaluation that could result from any policy of the others. For choosing B, this evaluation is simply $v(1, 0, 1, 1, \dots)$, while, for choosing A, the evaluation can become quite complex. Instead of following this line here, we will use a similar idea when discussing the concept of responsibility in the next section, where we will discuss other criteria than rationality and social preferences.

4. Solutions Based on Other Ethical Principles and Sustainability Paradigms

4.1. Responsibility

Rather than asking what combinations of uncertain welfare levels we should prefer for future generations, one can also ask what responsibility we have regarding future welfare. We will sketch here a certain simple theory of responsibility designed to be applicable to problems involving multiple agents, uncertainty, and potential ethically-undesired outcomes (EUOs), as in our TE. We distinguish two major types of responsibility, *forward*- and *backward*-looking responsibility, the latter having two subtypes, *factual* and *counterfactual* responsibility. While forward-looking responsibility is about still-existing possibilities, an agent or group of agents has to reduce the probability of future EUOs ("responsibility *to*"), backward-looking responsibility ("responsibility *for*") is about past possibilities that would have reduced the probability of an EUO that actually occurred (factual responsibility, e.g., Nagel's unlucky drunken driver [27]) or could have occurred (counterfactual responsibility, e.g., Nagel's lucky drunken driver [27]). In all three types, the degree of responsibility is measured in terms of differences of probabilities of EUOs. Rather than giving a formal definition, it will suffice to discuss the details of this theory at the hand of several choices for what constitutes an EUO in our TE.

Let us start by considering that an EUO is simply a low welfare in generation 1. Then the degree of *forward*-looking respectively of G_0 is the absolute difference between the probability of low welfare in generation 1 when choosing A rather than B, which equals η. In other words, G_0 would have a degree of η responsibility to choose A in order to avoid the EUO that G_1 gets low welfare. If they choose B instead, they will have a degree of *factual* backward-looking responsibility for G_1's low welfare equaling again η since this is the amount by which they could have reduced the probability of the EUO. If they behave "responsibly" by choosing A, G_1's welfare might also be low (with probability π), but G_0 would still not have backward-looking responsibility since they could not have reduced that probability.

If the EUO was simply a low welfare in G_2 rather than G_1, the assessment of G_0's responsibility must consider the possible actions of G_1 in addition to those of G_0. If G_0 chooses B, the probability of the EUO is zero, while if they choose A, it depends on G_1's choice. If G_1 would choose B, the EUO has probability 1 so that G_0's choice would make a difference of 1, while if G_1 would choose A, the EUO has probability $1 - \eta^2 < 1$ and G_0's choice would make a difference of only $1 - \eta^2 < 1$. In both cases, however, they have considerable forward-looking responsibility to choose B since by that they can reduce the probability of the EUO significantly. If choosing B, no backward-looking respectively accrues. If G_0 and G_1 both choose A and the collapse occurs at time 2, G_1 has no factual responsibility since they could not have reduced that probability, but G_0 has factual responsibility of degree $1 - \eta^2$. If G_0 chooses A and G_1 B, G_1's factual responsibility is η as seen above, but G_0's is even larger, since in view of G_1's actual choice, G_0 could have reduced the probability from one to zero by choosing B instead. Thus, G_0 has factual responsibility of 1. It might seem counterintuitive at first that the sum of the factual responsibilities of the two agents regarding that single outcome would be larger than 100%, but our theory is actually explicitly designed to produce this result in order to show that responsibility cannot simply be divided. Otherwise, each individual in a large group of bystanders at a fight in public could claim to have almost no responsibility to intervene (diffusion of responsibility). Finally, if both G_0 and G_1 choose B and no collapse happens, G_0 still has *counterfactual* responsibility since the collapse could have happened and G_0 could have reduced that probability by $1 - \eta^2$. This distinction between factual and counterfactual responsibility would also allow a discussion of Nagel's concept of moral luck in consequences [27] and responses to it, such as [28] but we will not go there here.

If the EUO is low welfare in G_3, it becomes more complicated. By choosing B, G_0 can avoid the EUO for sure, but when choosing A they might hope G_1 will choose B and the EUO will be avoided for sure as well, in which case they might claim to have a rather low responsibility to choose B which amounts only to π, the probability that G_1 will have no chance of choosing B due to immediate collapse. Common sense, however, shows that while wishful thinking regarding the actions of others might affect one's own psychological assessment of responsibilities, it cannot be the basis for an ethical observer's assessment of responsibility. Otherwise, even in a group of just two bystanders, neither one would be ethically obliged to intervene since both could hope the other does. Here we even take the opposite view and argue that G_0's degree of forward-looking responsibility should equal the *largest* possible amount by which they might be able to reduce the probability of the EUO, maximized over all possible behaviors of the other agents. This means that rather than being optimistic about G_1's action, they need to be pessimistic about both G_1's and G_2's behavior. The worst that can happen regarding the welfare of G_3 when G_0 chooses A is that G_1 would choose A and G_2, B. In that case, the EUO has probability 1, so G_0 would still be fully responsible (degree 1) to choose B in order to avoid the EUO.

Now what definition of EUO should we actually adopt in our TE? Two candidates seem natural, either a low welfare in any generation should already constitute an EUO (in which case it cannot be avoided by either A or B), or only an infinite number of low welfare generations, i.e., an eventual collapse into state T, should constitute an EUO. In the latter case, each generation in L has 100% forward-looking responsibility to choose B, and if they choose A instead, they will end up having 100% factual responsibility for the eventual collapse, regardless of the choices of later generations. Summarizing, we argue that a theory of responsibility that avoids the diffusion of responsibility and wishful thinking will deem B the responsible action in our TE since it avoids the worst for sure, even though this makes G_0 responsible for G_1's suffering.

4.2. Safe Operating Space for Humanity

In the following we continue our analyses of the ethical aspects of the TE from the perspective of the safe operating space (SOS) for humanity [2]. The SOS is located within planetary boundaries (PBs) "with respect to the Earth system" which "are associated with the planet's biophysical subsystems and or processes" [2]. The SOS is a fairly new concept for environmental governance, encapsulating several established concepts, such as the limits to growth [29,30], safe minimum standards [31–33], the precautionary principle [34], and the tolerable windows concept [35,36]. We let our analysis guide by the three "main" articles around the planetary boundaries and the SOS concepts [2–4], which have, at the time of this writing, together well over ten thousand citations, so that a comprehensive review of the SOS debate is beyond the possibilities of this article. We will, therefore, incorporate other papers only selectively.

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

One main difference to the approaches covered in the previous sections is the level of mathematical formalization. While we do acknowledge that some attempts of mathematical formalization of a SOS decision paradigm have been made [37], the original and most of the subsequent works do not provide a mathematical operationalization.

First of all we assess whether our TE is a suitable model within which the SOS concept can be applied at all. Rockström et al. [3] acknowledges that "anthropogenic pressures on the Earth System have reached a scale where abrupt global environmental change can no longer be excluded", which "can lead to the unexpected crossing of thresholds that drive the Earth System, or significant sub-systems, abruptly into states deleterious or even catastrophic to human well-being". Therefore, the authors "propose a new approach to global sustainability in which we define planetary boundaries within which we expect that humanity can operate safely." These lines resemble very well the situation in our TE where the decision-maker faces either a transition from L to T or from L to S.

However, the authors of the three papers in question do not mention any unfavorable P-like states on the way from L to S. Rockström et al. [2] states that "the evidence so far suggests that, as long as the thresholds are not crossed, humanity has the freedom to pursue long-term social and economic development." Emphasizing the long-term aspect, the last quote at least does not exclude the possibility of unfavorable interim states P on the way to safe, long-term "shelter" states S.

Nevertheless, opposing to the view that the SOS can be applied to the decision problem in our TE, the planetary boundaries' "precautionary approach is based on the maintenance of a Holocene-like state of the ES [Earth System]" [4]. This is emphasized because the "thresholds in key Earth System processes exist irrespective of peoples' preferences, values or compromises based on political and socioeconomic feasibility, such as expectations of technological breakthroughs and fluctuations in economic growth." [3]. One could argue that a mere transition from state L to S has to be interpreted as "destabilizing" [4]. However, this view disregards that our TE does not tell anything about the Holocene-likeness of the states L, T, P, and S. One may very well interpret states L, P, and S as Holocene-like. Further, as stated above, the ultimate justification for the planetary boundaries is to avoid Earth system states "catastrophic to human well-being" [3]. It is the only precautionary principle used by the PB approach that suggests staying within Holocene-like state.

Another opposition to the view that the SOS can be applied to the decision problem in our TE may result from the fact that "the planetary boundaries approach as of yet focuses on boundary definitions only and not as a design tool of compatible action strategies" [3]. The "PB framework as currently construed provides no guidance as to how [. . .] the maintenance of a Holocene-like state [. . .] may be achieved [. . .] and it cannot readily be used to make choices between pathways for piecemeal maneuvering within the SOS or more radical shifts of global governance" [4]. We make two observations from these quotes: First, the PB framework may not be used to guide how Holocene-like states shall be maintained, but it can surely be used as a guiding principle that Holocene-like states shall be maintained. Second, these quotes suggest that the authors assume that we are still currently in a Holocene-like SOS, since they do not explicitly account for re-entering it. However, one of the key messages of all three papers is that humanity has already crossed several of the nine planetary boundaries. One could conclude that humanity has, therefore, left the SOS.

The ultimate question regarding our TE is which states of our TE correspond to the SOS. Interpreting the T state as the catastrophic state that is to be avoided, four options seem plausible to constitute the SOS: (i) S; (ii) P and S; (iii) L and S; (iv) L, P, and S. State S is clearly part of the SOS. As mentioned above, the three papers avoid discussing P-like states. Therefore, both possibilities must be considered: either P-like states belong to the SOS or they do not. Regarding whether state L belongs to the SOS, [2] states: "Determining a safe distance [from the thresholds] involves normative judgments of how societies choose to deal with risk and uncertainty". This clearly reflects the circumstance that real-world environmental governance always has to account for risks and uncertainties. However, also in our TE we can associate the "risk" with the probability π of transitioning to state T under action A. Thus, if our decision-maker judges the risk π to be acceptable, L belongs to the SOS.

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

What are the consequences of assuming the SOS is composed of either of the sets (i)–(iv)? (i) If only S belongs to the SOS, one should choose action B, take the suffering of the next generation into account and finally end up in the SOS. There "humanity [can] pursue long-term social and economic development" [2]; (ii) If P, but not L, belongs to the SOS, the decision is still to take action B since that moves them even faster into the SOS; (iii) If L, but not P, belong to the SOS and we interpret the transition L → P → S as a "radical shift [. . .] of global governance" [4], the SOS concept "cannot [. . .] be used to make choices between pathways" [4], i.e., would be of no help here. Denoting that transition as "radical" can be justified since it temporarily leaves the SOS; Finally, (iv) assuming all of L, P, and S belong to the Holocene-like SOS, the SOS concept still "cannot readily be used to make choices between pathways for piecemeal maneuvering within the safe operating space" [4].

Overall, we conclude that whether or not the initial state L belongs to the SOS is essential for whether the SOS concept can be used to guide decisions in our TE. If L does not belong to the SOS, the decision problem is solved by taking action B. Otherwise the concept explicitly states that it cannot give guidance facing the trade-off highlighted in our TE.

4.3. Sustainability Paradigms à la Schellnhuber

Schellnhuber [38,39] proposes a set of five sustainability paradigms as idealizations of decision principles for governing the co-evolutionary dynamics of human societies and the environment as a part of a broader control-theoretical framework for Earth system analysis (also referred to as *geocybernetics*). The framework is introduced for deterministic systems and does not explicitly accommodate for probabilistic dynamics in the original publications, although it can be generalized to that case (as will be necessary in some of the interpretations of the sustainability paradigms for the TE given below). It also assumes that each *co-state* of the system under study consists of societal and environmental dimensions. In the context of our TE, the societal dimension corresponds to the welfare associated to a state. Since the TE does not explicitly specify evaluations of the environmental dimension, we assume here that it is mainly in line with the societal dimension, i.e., that it is "good" in states L and S and "bad" in state T. Regarding state P, we will discuss both possibilities below. The precise nature of this assignment does not impact most of the conclusions drawn below. In the following, we discuss the implications of the sustainability paradigms of *standardization*, *optimization*, *pessimization*, *equitization*, and *stabilization* introduced in [38] for our TE and relate them to the principles evaluated above.

4.3.1. Standardization

When adhering to the standardization paradigm, decisions on actions follow prescribed "environment and development" standards based on upper or lower limits on various system variables or aggregated indicators. The standardization paradigm includes governance frameworks such as the tolerable windows approach [36], climate guardrails and planetary boundaries [2,3] (see also Section 4.2). Following a pure standardization paradigm may lead to problematic and unintended outcomes, since system dynamics are not explicitly taken into account.

Several examples for concrete flavors of the standardization paradigm are of interest in analyzing the TE. In the case of *eco-centrism,* only environmental standards are taken into account (requiring a "good" environmental state for all time). If the environment is assumed to be in a good condition in state P, then clearly following this eco-centric paradigm implies choosing action B. However, if state P is interpreted as bad for the environment, then the eco-centric paradigm seems to imply choosing A to conserve the local environment at least with probability η rather than degrading it for certain, temporarily. In the case of a *tolerable environment and development window,* both societal and environmental dimensions are taken into account (requiring a good environmental state and a high societal welfare for all time). This variant of the standardization paradigm does not allow reaching a decision on which action to choose, because both actions A and B violate the standards at some point. A third example for a standardization paradigm is the *maintenance of living standards*: for all times a certain level of minimum wealth should be maintained (living standard may be measured by more complex aggregated indicators in higher-dimensional models). A short-sighted society would choose action A following this paradigm since the standard is fulfilled with probability η per generation. Adopting a second-best interpretation requiring the standard to be met only after some time, a more farsighted society would choose action B, meeting the standard when reaching a state with certainty S in generation 2.

4.3.2. Optimization

The optimization paradigm is based on "wanting the best" [38] and selects actions accordingly to maximize a given utility function. It is, hence, closely related to the rational choice framework and its implications for the TE discussed in Section 3. Optimization can be performed under constraints given by standards, resulting in a combination of the optimization and standardization paradigms. As seen already in Section 3, adopting the optimization paradigm carries a risk related to the considerable uncertainty on whether future generations will actually be willing or able to follow the previously-determined optimal management sequence.

4.3.3. Pessimization

The pessimization paradigm is based on the principle of "avoiding the worst" and is, hence, also referred to as an "Anti-Murphy strategy of sustainable development" [38]. It is a resilience-centered paradigm that calls for excluding management sequences that could allow for disastrous mismanagement by future generations. An example for a specific pessimization paradigm is the minimax strategy that dictates to minimize the maximum possible damage caused by a management sequence. The rationale is, hence, to hedge the damage that can be done by the management choices of future generations. With respect to the TE, this calls for choosing action B to avoid the worst outcome: to likely get trapped in the degraded state T forever caused by future generations repeatedly choosing action A.

4.3.4. Equitization

The equitization paradigm is centered around avoiding inequalities of various kinds, be it geographical or temporal. Focusing on the second aspect of inter-generational equity here, it describes a quest for just allocation of choices in time to keep the space of management options open for future generations. Extending upon the Brundtland definition of sustainable development focusing on being able to meet the needs (welfare) of the current and future generations, the equitization paradigm demands the "equality of environment and development options for successive global generations" [38]. Since open and fairly distributed option spaces are key for allowing future generations to adapt and transform to deal with previously unknown and unforeseeable perturbations and challenges, the equitization paradigm is closely related to principles from resilience thinking. If we interpret the choice between A and B in our TE as a kind of "development option" in the sense of Schellnhuber, then the equitization paradigm seems to call for choosing A, since this preserves options for the next generation with at least probability η. For option B, the generation in P and future generations in S would have no options left after all. It is interesting to note that in our deliberately simple and fully-known system described in the TE, following the equitization paradigm would, therefore, keep the system in the risky state L and would not allow navigating to the desirable state S. On the other hand, if we rather interpret "development options" as an aspect of high welfare, clearly state S provides more options than T, so we would be back to the question of whether one should sacrifice the options of one generation for the options of all later generations.

4.3.5. Stabilization

The stabilization paradigm describes the goal of steering the system towards a preselected state or set of states that is considered sustainable. For example, it encapsulates the underlying intentions

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

of the United Nations Sustainable Development Goals [40], and other political agreements of that type, to inform and steer governance for sustainable development. In the TE, the stabilization clearly paradigm demands to choose action B, since only then the desirable state S can be reached where high wealth can be sustained for all time.

5. Discussion

When designing the thought experiment discussed in this article, the authors originally had the intuition that most schools of thought would provide a relatively clear answer to the seemingly simple question of whether the hypothetical generation finding themselves in situation L should choose option A or B. Indeed, many individuals we discussed it with seemed to have a strong immediate gut feeling as to what one "should" do in that situation. For example, when one of the authors asked two practicing Buddhists, who have discussed the Buddhist worldview with each other for years, about their opinion, both immediately announced the Buddhist position on this would be perfectly clear. For the formally slightly similar trolley problem, a survey among professional philosophers showed that only 24% of respondents would not take a position on that problem [41].

However, as it turns out in light of the above analyses, we could not have been more mistaken. When asked to explain, the two Buddhists mentioned above argued very convincingly from their respective interpretation of Buddhism, one for action A, but the other for action B. We had similar experiences with people adhering to the schools of thought we chose to discuss in this article. As the above analysis shows, neither the optimal control framework, welfare economics, game theory, the concept of a safe operating space, or many of the discussed sustainability paradigms give a really clear and unambiguous answer to the question, at least not without having to choose parameters such as the right time horizon, level of inequality aversion, risk attitude, preference for fairness, etc. In some cases, the ambiguity also seems to be due to difficulties in matching the terminology and basic concepts of a framework for evaluation to the situation described in the TE. Even seemingly clear concepts such as "options", "inequality", "risk", etc. become complicated to apply and assess when they are entangled in the way they are in our TE.

Overall, our impression is that much of the difficulties have to do with the strong presence of probabilistic uncertainties and their strong correlations over time caused by the extreme form of lock-in effects in our TE. Once choosing action B or once collapsing into state T, there is no turning back, and some of our analysis depends on this extreme assumption. While the assumption might be criticized as unrealistic, there is no denying that, also in the real world, choices such as a transition to a decarbonized economy or events such as the GHG-emissions-induced tipping of a climatic tipping element will have very long-lasting effects which, for the sake of an evaluation, might just as well be assumed to be effectively irreversible. Still, future work on this and similar thought experiments should also assess whether certain modifications, such as (i) the introduction of a small probability of being able to return to state L from either T or S; (ii) exogenous or endogenous changes in the definition of "welfare" over time; or (iii) status effects, such as anticipated posterior perception ("making history"), to name only a few, would make a qualitative difference.

The presence of strong uncertainties is less debatable than that of irreversible lock-ins, thus, it is somewhat surprising that when trying to apply modern concepts, such as some of the sustainability paradigms discussed in Section 4, it seems that they are not really made for choice situations where consequences involve high and long-lasting uncertainties, unclear causal relationships, and the possible necessity of temporary reductions in welfare. In particular, regarding the latter aspect, our impression is that discussing intermediate suffering is somewhat unpopular in the sustainability discourse. Since potential trade-offs between intermediate suffering and long-term sustained welfare might exist not only in our TE, but also in the real world, this calls for a debate among scholars and policy-makers of how to handle this trade-off.

Still, we argue that a few patterns of evaluation emerged quite clearly across the different schools of thought. Most prominently, but least surprisingly, a focus on the farther future and the long-term evolution clearly makes option B more attractive than A. Second, a strong preference for equality across generations, whether expressed via a large coefficient of inequality aversion in a rationality-based framework, or by choosing to follow the equitization paradigm, seems to make option A more attractive overall since it distributes welfare, options, and risks more evenly over time. Readers who perceive these results as rather unsurprising will hopefully consider them a kind of sanity check for our setup. In addition to these intuitive results, there were also a few surprises, including the rather easy occurrence of time-inconsistencies or probabilistic elements in optimal policies even in the single-agent interpretation, caused by inequality aversion, or the occurrence of alternating recommendations to choose A or B in consecutive generations in the multi-agent interpretation. The most interesting result of our study, however, is probably the overall insight that even such a simple and seemingly clear setup as the TE presented here can generate such a diverse and complex set of assessments even within a single well-established framework, such as the welfare-function-based one. While the flexibility of the welfare function approach due to its many possible specifications and continuous parameters may be considered its main weakness, we believe there still remains to be found a convincing basic ethical principle that would make a clearer recommendation and can be hoped to be accepted as overriding all other approaches.

We, therefore, close with a few suggestions as to which additional approaches and which modifications of the TE might be promising. Adding a clearer quantitative distinction between the welfare levels in states L, T, P, and S might resolve certain ties in the welfare framework, but might also distract from the basic qualitative problem by focusing too much on quantities. If one would identify option A as the "default", or rather "passive" choice, and B as more "active", one could apply concepts such as the Doctrine of the Double Effect [42], which have been used to study the trolley problem and similar dilemmas. This, and similar additional details in the description of the TE might also allow an assessment in terms of religious traditions and other moral codes.

Author Contributions: J.H. conceived and designed the thought experiment, and all authors performed the analysis and wrote the paper.

Acknowledgments: This work was conducted in the framework of the project on Coevolutionary Pathways in the Earth System (COPAN) at the Potsdam Institute for Climate Impact Research (PIK). The authors wish to thank three anonymous referees for helpful comments, and thank Maddalena Ferranna, Marc Fleurbaey, Ulrike Kornek, Adrian Lison, and the COPAN project team at PIK for intensive discussions and some references. We are grateful for financial support by the Stordalen Foundation (via the Planetary Boundary Research Network PB.net), the Earth League's EarthDoc Programme, Leibniz Association (project DOMINOES), and the Heinrich Böll Foundation.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Zalasiewicz, J.; Williams, M.; Haywood, A.; Ellis, M. The Anthropocene: A New Epoch of Geological Time? *Philos. Trans. A Math. Phys. Eng. Sci.* **2011**, *369*, 835–841. [CrossRef] [PubMed]
2. Rockström, J.; Steffen, W.; Noone, K.; Persson, A.; Chapin, F.S.; Lambin, E.F.; Lenton, T.M.; Scheffer, M.; Folke, C.; Schellnhuber, H.J.; et al. A Safe Operating Space for Humanity. *Nature* **2009**, *461*, 472–475. [CrossRef] [PubMed]
3. Rockström, J.; Steffen, W.; Noone, K.; Persson, A. Planetary Boundaries: Exploring the Safe Operating Space for Humanity. *Ecol. Soc.* **2009**, *14*, 32. [CrossRef]
4. Steffen, W.; Richardson, K.; Rockstrom, J.; Cornell, S.E.; Fetzer, I.; Bennett, E.M.; Biggs, R.; Carpenter, S.R.; de Vries, W.; de Wit, C.A.; et al. Planetary Boundaries: Guiding Human Development on a Changing Planet. *Science* **2015**, *347*, 1259855. [CrossRef] [PubMed]
5. Raworth, K. A Safe and Just Space For Humanity: Can We Live within the Doughnut? *Oxfam Policy Pract. Clim. Chang. Resil.* **2012**, *8*, 1–26.
6. Lenton, T.M.; Held, H.; Kriegler, E.; Hall, J.W.; Lucht, W.; Rahmstorf, S.; Schellnhuber, H.J. Tipping Elements in the Earth's Climate System. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 1786–1793. [CrossRef] [PubMed]

7. Schellnhuber, H.J. Tipping Elements in the Earth System. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 20561–20563. [CrossRef] [PubMed]
8. Rockström, J.; Gaffney, O.; Rogelj, J.; Meinshausen, M.; Nakicenovic, N.; Schellnhuber, H.J. A Roadmap for Rapid Decarbonization. *Science* **2017**, *355*, 1269–1271. [CrossRef] [PubMed]
9. Schellnhuber, H.J. Geoengineering: The Good, the MAD, and the Sensible. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 20277–20278. [CrossRef] [PubMed]
10. Vaughan, N.E.; Lenton, T.M. A Review of Climate Geoengineering Proposals. *Clim. Chang.* **2011**, *109*, 745–790. [CrossRef]
11. Kleidon, A.; Renner, M. A Simple Explanation for the Sensitivity of the Hydrologic Cycle to Surface Temperature and Solar Radiation and Its Implications for Global Climate Change. *Earth Syst. Dyn.* **2013**, *4*, 455–465. [CrossRef]
12. Kopp, R.E.; Shwom, R.; Wagner, G.; Yuan, J. Tipping Elements and Climate-Economic Shocks: Pathways toward Integrated Assessment. *Earth's Future* **2016**, *4*, 362–372. [CrossRef]
13. Donges, J.F.; Winkelmann, R.; Lucht, W.; Cornell, S.E.; Dyke, J.G.; Rockström, J.; Heitzig, J.; Schellnhuber, H.J. Closing the Loop: Reconnecting Human Dynamics to Earth System Science. *Anthr. Rev.* **2017**, *4*, 151–157. [CrossRef]
14. Van Vuuren, D.P.; Lucas, P.L.; Häyhä, T.; Cornell, S.E.; Stafford-Smith, M. Horses for Courses: Analytical Tools to Explore Planetary Boundaries. *Earth Syst. Dyn.* **2016**, *7*, 267–279. [CrossRef]
15. Heitzig, J.; Donges, J.F.; Barfuss, W.; Kassel, J.A.; Kittel, T.; Kolb, J.J.; Kolster, T.; Müller-Hansen, F.; Otto, I.M.; Wiedermann, M.; et al. Earth System Modelling with Complex Dynamic Human Societies: The copan:CORE World-Earth Modeling Framework. *Earth Syst. Dyn. Discuss.* **2018**, 1–27. [CrossRef]
16. Horowitz, T.; Massey, G.J. *Thought Experiments in Science and Philosophy*; Rowman & Littlefield Publishers: Lanham, MD, USA, 1991.
17. Heitzig, J.; Kittel, T.; Donges, J.F.; Molkenthin, N. Topology of Sustainable Management of Dynamical Systems with Desirable States: From Defining Planetary Boundaries to Safe Operating Spaces in the Earth System. *Earth Syst. Dyn.* **2016**, *7*, 1–30. [CrossRef]
18. Kittel, T.; Koch, R.; Heitzig, J.; Deffuant, G.; Mathias, J.-D.; Kurths, J. Operationalization of Topology of Sustainable Management to Estimate Qualitatively Different Regions in State Space. *arXiv*, **2017**, arXiv:1706.04542v3.
19. Hansson, S.O. Social Choice with Procedural Preferences. *Soc. Choice Welf.* **1996**, *13*, 215–230. [CrossRef]
20. Fleurbaey, M. On Sustainability and Social Welfare. *J. Environ. Econ. Manag.* **2015**, *71*, 34–53. [CrossRef]
21. Chichilnisky, G. An Axiomatic Approach to Sustainable Development. *Soc. Choice Welf.* **1996**, *13*, 231–257. [CrossRef]
22. Sozou, P.D. On Hyperbolic Discounting and Uncertain Hazard Rates. *Proc. R. Soc. B Biol. Sci.* **1998**, *265*, 2015–2020. [CrossRef]
23. Barberis, N. Thirty Years of Prospect Theory in Economics: A Review and Assessment. *J. Econ. Perspect.* **2013**, *27*, 173–196. [CrossRef]
24. Sen, A. Informational Bases of Alternative Welfare Approaches. *J. Public Econ.* **1974**, *3*, 387–403. [CrossRef]
25. Epstein, L.G.; Zin, S.E. Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework. *Econometrica* **1989**, *57*, 937–969. [CrossRef]
26. Von Neumann, J.; Morgenstern, O. *Theory of Games and Economic Behavior (Commemorative Edition)*; Princeton University Press: Princeton, NJ, USA, 2007.
27. Nagel, T. Moral Luck. In *Mortal Questions*; Cambridge University Press: Cambridge, UK, 1979.
28. Andre, J. Nagel, Williams, and Moral Luck. *Analysis* **1983**, *43*, 202–207. [CrossRef]
29. Meadows, D.H.; Meadows, D.L.; Randers, J.; Behrens, W.W. The Limits to Growth. *N. Y.* **1972**, *102*, 27.
30. Meadows, D.; Randers, J.; Meadows, D. *A Synopsis: Limits to Growth: The 30-Year Update*; Chelsea Green Publ. Co.: White River Junction, VT, USA, 2004.
31. Ciriacy-Wantrup, S.V. *Resource Conservation: Economics and Policies*; University of California Press: Berkeley, CA, USA, 1963.
32. Bishop, R.C. Endangered Species and Uncertainty: The Economics of a Safe Minimum Standard. *Am. J. Agric. Econ.* **1978**, *60*, 10–18. [CrossRef]
33. Crowards, T.M. Safe Minimum Standards: Costs and Opportunities. *Ecol. Econ.* **1998**, *25*, 303–314. [CrossRef]

34. Raffensperger, C.; Tickner, J.A. *Protecting Public Health and the Environment: Implementing the Precautionary Principle*; Island Press: Washington, DC, USA, 1999.
35. Zimmermann, H.; Schellnhuber, H.-J. *Scenario for the Derivation of Global CO_2 Reduction Targets and Implementation Strategies. Statement on the Occasion of the First Conference of the Parties to the Framework Convention on Climate Change in Berlin*; WBGU: Bremerhaven, Germany, 1995.
36. Petschel-Held, G.; Block, A.; Cassel-Gintz, M.; Kropp, J.; Lüdeke, M.K.B.; Moldenhauer, O.; Reusswig, F.; Schellnhuber, H.-J. Syndromes of Global Change: A Qualitative Modelling Approach to Assist Global Environmental Management. *Environ. Model. Assess.* **1999**, *4*, 295–314. [CrossRef]
37. Barfuss, W.; Donges, J.F.; Lade, S.; Kurths, J. When Optimization for Governing Human-Environment Tipping Elements Is Neither Sustainable nor Safe. *Nat. Commun.* **2018**; in print.
38. Schellnhuber, H.J. Discourse: Earth System Analysis—The Scope of the Challenge. *Earth Syst. Anal. Integr. Sci. Sustain.* **1998**, 3–195. [CrossRef]
39. Schellnhuber, H.-J. 'Earth System' Analysis and the Second Copernican Revolution. *Nature* **1999**, *402*, C19–C23. [CrossRef]
40. *Transforming Our World: The 2030 Agenda for Sustainable Development*; United Nations Publishing: Herndon, VA, USA, 2015.
41. Bourget, D.; Chalmers, D.J. What Do Philosophers Believe. *Philos. Stud.* **2014**, *170*, 465–500. [CrossRef]
42. Foot, P. The Doctrine of Double Effect. *Oxf. Rev.* **1967**, *5*, 5–15.

Reproduced from: J. Heitzig et al., A thought experiment on sustainable management of the earth system, Sustainability, vol. 10, no. 6, pp. 1-25, 2018, doi: 10.3390/su10061947. Published under Creative Commons Attribution License 4.0 (CC BY).

3.3 *Dynamics of adaptive social-ecological networks*

IN THIS THIRD SECTION we present investigations of social-ecological systems represented as complex networks [Lade et al., 2017].

The first paper, "The Dynamics of Coalition Formation on Complex Networks" [Auer et al., 2015], is dedicated to analyse the formation of self-organizing domains of cooperation ("coalitions") on an acquaintance network.

Another relevant social dynamics on networks is the spreading of opinions, behaviours, decision making or social norms, coevolving with processes such as homophily that change the underlying network structure. A prominent model is the so-called "adaptive voter model" [Holme and Newman, 2006], used in numerous studies, for example investigating zealotry effects [Klamser et al., 2017]. In "Macroscopic description of complex adaptive networks coevolving with dynamic node states" [Wiedermann et al., 2015] we studied decision making on local resource use and social learning on an adaptive complex network which is dependent on the individual dynamic state of each node and vice versa. This serves as a stylised representation, the copan:EXPLOIT model, of the coevolution of renewable resources with the dynamics of harvesting agents on a social network.

We close this section with "The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources" [Geier et al., 2019]. In this third exemplary modelling approach, we extended the renewable resource layer and the resource user layer in the network from the previous contribution by a third layer representing governance agents that can penalise unsustainable resource use. We uncovered that a sustainable outcome depends on parameters of each network layer and observed interesting resilience trade-offs between an "eco-dictatorship" and adaptive polycentric governance by multiple individual actors.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

The Dynamics of Coalition Formation on Complex Networks

S. Auer[1,2], J. Heitzig[2], U. Kornek[2], E. Schöll[1] & J. Kurths[2,3,4,5,6]

Received: 17 June 2015
Accepted: 22 July 2015
Published: 25 August 2015

Complex networks describe the structure of many socio-economic systems. However, in studies of decision-making processes the evolution of the underlying social relations are disregarded. In this report, we aim to understand the formation of self-organizing domains of cooperation ("coalitions") on an acquaintance network. We include both the network's influence on the formation of coalitions and vice versa how the network adapts to the current coalition structure, thus forming a social feedback loop. We increase complexity from simple opinion adaptation processes studied in earlier research to more complex decision-making determined by costs and benefits, and from bilateral to multilateral cooperation. We show how phase transitions emerge from such coevolutionary dynamics, which can be interpreted as processes of great transformations. If the network adaptation rate is high, the social dynamics prevent the formation of a grand coalition and therefore full cooperation. We find some empirical support for our main results: Our model develops a bimodal coalition size distribution over time similar to those found in social structures. Our detection and distinguishing of phase transitions may be exemplary for other models of socio-economic systems with low agent numbers and therefore strong finite-size effects.

Statistical physics provides a powerful means to conceptually study mechanisms of socio-economic systems and their associated transformations such as market restructuring, social upheavals and revolutions. Many socio-economic systems exhibit network structures[1], and a number of studies show how network structures influence behaviour such as bilateral cooperation[2]. Much less work is done on the reverse effect that the network structure in turn adapts to behaviour[3–5]. While both processes are interesting in themselves, in the context of opinion dynamics it is actually the feedback loop of both network adaptation and dynamics on the network which leads to the most interesting nonlinear effects. E.g., the seminal work of Holme[6] presents a model in which a phase transition occurs that can be interpreted as a great transformation.

In this report, we transfer the methods of Holme[6] from local social dynamics to a more complex form of mesoscopic social self-organization, namely that of multilateral cooperation (here called *coalitions*), whose interaction with network structures has not been studied before. In particular, we present a model of the coevolution of an adaptive network representing social acquaintance and a coalition structure which is a partition of nodes into coalitions of arbitrary size representing multilateral cooperation (Fig. 1). Instead of an exogenously given number of opinion groups as previously studied in the literature, the number of coalitions in our model evolves endogenosly as a process of self-organization from the boundedly rational behavior of the agents. Our model can be applied to socio-economic environments where cooperation promises economic or social advantages, and to study such diverse subjects as firm size distributions, fish cohorts, and political parties. Our methods to detect phase transitions are especially applicable to small real-world systems. However, in our case low sample sizes do not necessarily

[1]Institute for Theoretical Physics, Technische Universität Berlin—Hardenbergstr. 36, 10623 Berlin, Germany, EU. [2]Potsdam Institute for Climate Impact Research—P.O. Box 60 12 03, 14412 Potsdam, Germany, EU. [3]Department of Physics, Humboldt University—Newtonstr. 15, 12489 Berlin, Germany, EU. [4]Institute for Complex Systems and Mathematical Biology, University of Aberdeen—Aberdeen AB24 3FX, UK, EU. [5]Department of Control Theory, Nizhny Novgorod State University—Gagarin Avenue 23, 606950 Nizhny Novgorod, Russia. [6]Institute of Applied Physics of the Russian Academy of Sciences—Uljanov str. 46, 603950 Nizhny Novgorod, Russia. Correspondence and requests for materials should be addressed to S.A. (email: auer@pik-potsdam.de)

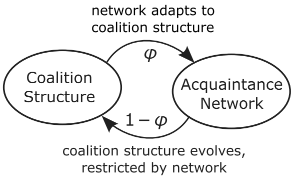


Figure 1. Scheme of coevolution. In each period, either one random cross-coalition links is replaced by an intra-coalitional link (adaptation with rate ϕ) or some random agent changes the coalition structure (coalition formation with rate $1-\phi$), where each two members of a coalition must be connected by a path in the network.

indicate interaction processes with a low number of people since each agent might be composed of many individuals, already. A common economic situation for which cooperation is critical is the use of a common pool resource. It leads to nontrivial coalition formation dynamics because agents not only have an incentive to form a coalition but also to leave a coalition in order to profit from the efforts of the remaining coalition. Since one of the major current economic challenges, the transition to a low-carbon economy, is closely related to several common pool resources like the atmosphere and renewable energy sources, we focus on the application of our model to common pool resources in this article.

Results

In our model, each coalition rationally decides how much of the resource to exploit and gets a corresponding payoff that depends on all coalitions' sizes and decisions. On this basis, individual agents rationally decide to form new coalitions with their acquaintances or merge or leave existing coalitions. Finally, they may also form new acquaintance links to members of their own coalition or break existing ones to members of other coalitions. The main control parameter in our model is the relative speed of acquaintance adaptation vs coalition formation, the adaptation rate ϕ, and the main feature of the resulting dynamics is the distribution of coalition sizes that evolves as an equilibrium over time.

For the case of agents exploiting a common pool resource[7], we find a second order phase transition when adaptation versus coalition formation crosses its critical value, $\phi = \phi_c$. For subcritical adaptation rates (see Methods for the description of the model and parameters), the coalition structure is dominated by very few macroscopic or even near-global coalitions. This leads to a peculiarly multimodal size distribution that can also be observed in various real-world systems[8–13], not only in socio-economic contexts but also in purely physical systems such as droplets[14]. In contrast, at the critical adaptation rate, a more heterogeneous but power-law-tailed size distribution with much smaller maximal coalitions emerges (see Figs 2 and 3).

Change in Coalition Size Distributions.

We see significant changes in the distribution from a few macroscopic coalitions to complex multimodal distributions, when the adaptation rate ϕ is changed. There are two extremes. For $\phi = 0$ the dynamics are purely based on coalition formation and hence coalition sizes approach the initial component sizes. In contrast, for $\phi = 1$ only network adaptation takes place: starting with a coalition structure of only singletons, this parameter setting immediately converges without any further changes taking place; from the beginning, there are no coalition partners to link with. For small ϕ the distribution has peculiar features: a linearly decreasing frequency for small coalition sizes s and one or two local maxima for larger coalition sizes. Its multimodal nature emerges endogenously from the nonlinear dynamics of coalition formation. As a matter of fact, from empirical observations, multimodal distributions of social structures are well known, e.g. multimodal size distributions have been found in growth patterns of fish cohorts[8,9] and droplet sizes[14], in human communication[10] and in firm and city size distributions of developing countries[11–13].

For a typical size of coalition forming systems, $N = 300$ nodes, a look at the distribution of s in steps of $\Delta\phi = 0.1$ initially does not reveal any interesting artifacts such as power-laws. However, at values of ϕ close to one, the local maxima at the tails disappear, and for a critical adaptation rate $\phi_c \approx 0.97$ coalition sizes show a power-law tail (Fig. 3c). The reason for such a high critical value are the more macroscopic effects of the coalition formation process as compared to the network adaptation process; it may involve hundreds of agents at once whereas network adaptation only affects three agents at a time (see Methods). In Fig. 4a, a higher resolution plot of maximum coalition size S vs. φ shows a turning point of this order parameter[15], something we expect for a second-order phase transition. It is not very distinctive but this is expected due to finite size effects[16]. Samples of up to 500 nodes are rather small for statistical physics and an exponential progression of system sizes would be more revealing as network distances scale slowly with N. Still, we chose a linear progression in N because the coalition formation process causes high computational costs with rising N and agent numbers of up to several hundreds are quite realistic for many socio-economic systems[17]. Nevertheless, phase transitions appear, only the accompanying singularities are washed out or smoothed[18] due to finite-size effects. Also, we expect only small finite-size

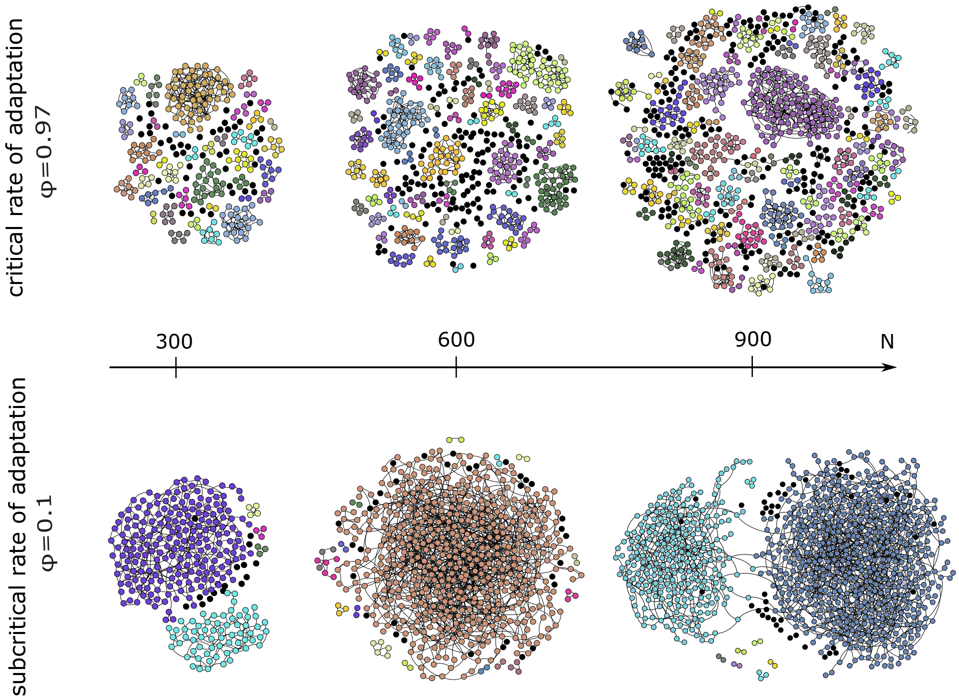


Figure 2. Acquaintance network with coalition structure (each color represents one coalition, black dots are singleton coalitions) for varying system size (columns: $N = 300$, $N = 600$ and $N = 900$) and adaptation rate (rows: $\phi = 0.97$ and $\phi = 0.1$). Note that some of the smaller network components consist of more than one coalition. Each network is the equilibrium result of one model run.

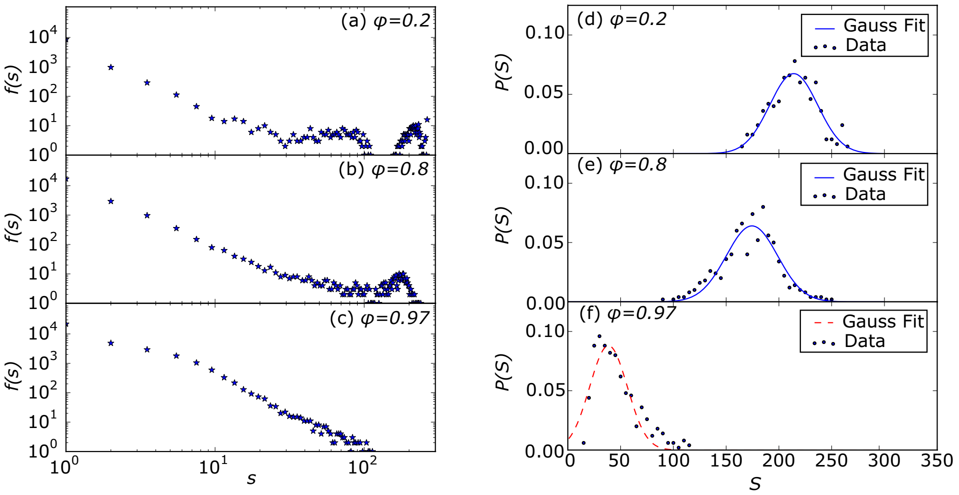


Figure 3. Left (**a**–**c**) log-log plot of frequency distribution of *all* coalition sizes s and right (**d**–**f**) histograms $P(S)$ of *maximum* coalition size S in the consensus state for $\phi = 0.2$, $\phi = 0.8$ and $\phi = 0.97$, respectively. $N = 300$ and $\langle k \rangle = 3$ (for 500 model runs).

corrections[19] to critical scaling with an exponential progression to larger N which are insignificant in the context of socio-economic modeling. Thus, the detailed analysis of finite-size phase transitions may be of great use for further models in this context.

Reproduced from: S. Auer et al., The Dynamics of Coalition Formation on Complex Networks, Sci. Rep., vol. 5, pp. 1–8, 2015, doi: 10.1038/srep13386. Published under Creative Commons Attribution License 4.0 (CC BY).

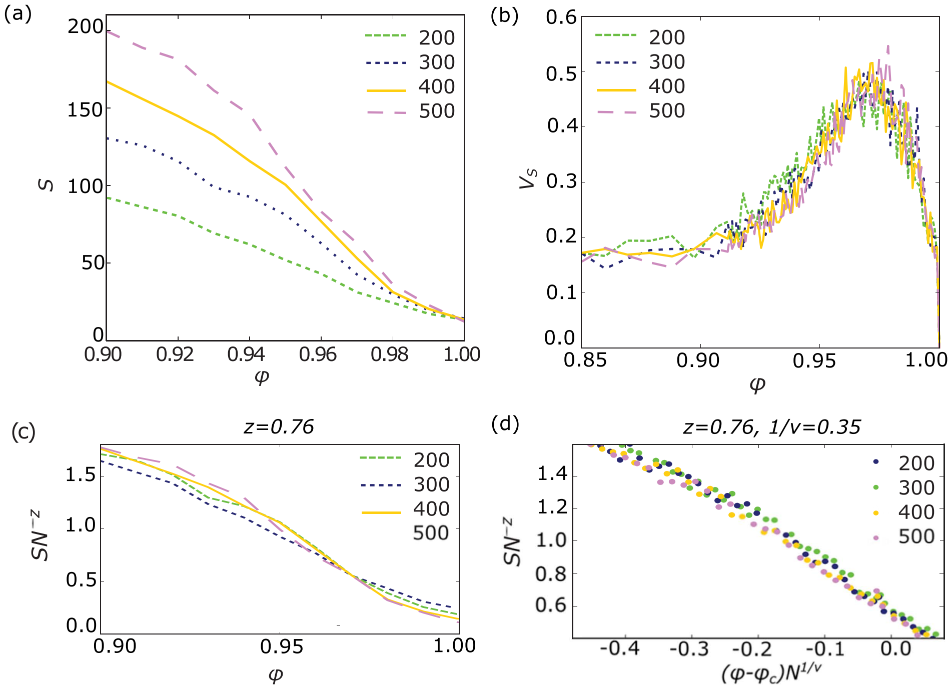


Figure 4. Plot of (**a**) order parameter *S*, (**b**) coefficient of variation V_S and (**c**) *S* scaled with N^{-z} over control parameter ϕ for different agent numbers *N*. (**d**) Data collapse close to the critical point ϕ_c. *S* scaled with N^{-z} over $(\phi - \phi_c)$ scaled with N^{ν}. All variables are averaged over 100 model runs.

Second-order Phase Transition. From these plots alone, the type of phase transition (first- or second-order) is hard to identify because a finite sample size will give both first- and second-order transitions a similar appearance. However, the type of transition is revealed by the probability density function of the order parameter *S*. For a varying control parameter ϕ, first-order transitions have two peaks at fixed position with changing height[20]. Whereas for a second-order transition, a Gaussian peak continuously changes its position—the maximum and mean values of *S* are moving to smaller values for increasing ϕ[18] until the Gaussian converts to a heavy-tailed function at the critical point. In our case, for $\phi = 0.2$ and $\phi = 0.8$, the Gaussian curve fits relatively well, but for $\phi = 0.97$ there is an obvious mismatch. For Fig. 3d with $\chi^2 = 21.7$, and for Fig. 3e with $\chi^2 = 29.6$, the Chi-squared test statistics is below the critical quantile, $\chi^2_{0.05,19} = 30.1$. This means that on the level 5% we cannot reject the hypothesis of a normal distribution. For Fig. 3f, $\chi^2 = 242.2$ exceeds the critical quantile $\chi^2_{0.05,13} = 22.4$. A long tail appears, featuring bigger coalition sizes that cannot be explained by a Gaussian distribution[18]. Both arguments underline the assumption of a continuous phase transition.

Quantification of the Scaling Relation. Via the maximum of the coefficient of variation of *S*, V_S, it is possible to identify the critical region (Fig. 4b). According to scaling theory, for different agent numbers, V_S should peak at about the same value of ϕ, only slightly shifted from the critical point by *const.* $\cdot N^{-1/\nu}$, where ν is a critical exponent[21]. In the region close to the expected critical point ϕ_c, we have estimated V_S for several agent numbers and indeed all maxima appear approximately at $\phi = 0.97$, in accordance with our earlier estimate. With this knowledge, it is possible to quantitatively grasp the critical dynamics. To determine the critical exponents, it is important to recall the classical scaling relation[21] and apply it to our model case, where ϕ takes the role of temperature and *S* the role of magnetization (see Methods for further explanation):

$$S = N^{-\frac{\beta}{\nu}} f\left((\phi - \phi_c) N^{\frac{1}{\nu}}\right), \tag{1}$$

where β is another critical exponent. At first, it should be possible to find a scaling exponent *z* for which SN^z intersects for all agent numbers in a single point $(\phi_c, f(0))$[22]. Therefore, we vary *z* until a value ϕ_c is found where all curves cross. This is the case for $z \approx 0.76$. At this point $z = \beta/\nu$. In Fig. 4c, the result of successfully scaling the order parameter is shown. After that, scaling $(\phi - \phi_c)$ by the factor $N^{\frac{1}{\nu}}$ will lead to a data collapse in a region closely drawn around the critical point ϕ_c[22]. This way, in Fig. 4d the critical exponent ν was found to be approximately $\nu = (0.35)^{-1}$.

Discussion

What can we infer from these results? If the acquaintance network in our coalition formation model adapts only slowly to the coalition structure, the formation of a grand coalition is most probable. Only for really high adaptivity, a fast transition to a heterogeneous coalition structure appears because then the effect of coalition formation is suppressed by a permanent rewiring of the acquaintance links. Before it is even possible to find a neighbor coalition to unite with, at some earlier stage the link to this coalition was already removed. If adaptivity represents some kind of social punishment (deprivation of social contact between agents from different coalitions) then in this case punishment would actually be counterproductive. At high frequency it leads to the isolation of a high number of small and midsize coalitions forming independent network components. However, full cooperation provides the highest benefits to all agents in many socio-economic situations, including the common pool context of our study. From an outside perspective, e.g. consumers facing an oligopoly, it may however be desirable to keep coalitions (cartels) small. Of course, this phenomenon of coalition isolation is caused by our assumption that the total number of acquaintances stays constant over time which has been argued to be approximately realistic in social relations[1]. As this assumption has such large effect on the model outcome and implications, it would be most interesting to study different scenarios in future work. In the context of contemporary issues, our findings can be used to support transformation processes by fostering the persistence of social networks by lowering ϕ. Both our model and real-world systems may undergo non-equilibrium phase transitions (in equilibrium physics an isolated system maximizes its entropy whereas non-equilibrium phase transition are driven by an external force, e.g. a heat bath, or control parameter[23]) and therefore the investigation of socio-economic transformations can profit from conceptual models of decision-making processes. Less realistic is the investigation of agents in a fixed state after model convergence. However, after only a few time steps we observe the same basic appearance of coalition size distributions. E.g. in the subcritical case, we see a multimodal distribution with the local maxima migrating to greater values with evolving time (similar to[14]). Still, this aspect is part of ongoing work.

In our model, we have increased the level of complexity from simple opinion adaptation processes from earlier research to more complex decision making determined by costs and benefits, and from local social interaction to mesoscopic cooperation. Our approach gets support from empirical data. In our model, the fat-tailed and bimodal coalition size distributions develop over time, they are model-inherent. The distributions resulting from such processes of self-organization deliver measurable quantities to study such transformation processes. Future work should vary the payoff sub-model in order to represent different archetypical socio-economic situations than the common-pool setting. One may then compare them with firm size distributions[12,13,17] from different economic sectors to identify the drivers of firm size growth. Other cases of multimodality in social systems were found in city size distributions of developing countries[11] and fish cohorts[8,9] possibly resulting from cooperative phenomena. However, even natural processes such as droplet growth give a similar picture: a power-law distribution for small droplets and a maximum for larger ones[14]. For the study of specific real-world systems, it might however be necessary to model heterogeneous agents. But most importantly, the observation of phase transitions with respect to network adaptivity in our model encourages follow-up work on the role that the relative speed of processes in social feedback loops has for transformation process.

Methods

We start with an Erdös-Rényi random acquaintance network and a coalition structure composed of one singleton "coalition" per agent (coalitions of size 1), representing no initial cooperation. Coalitions are collective decision-makers and the members of a coalition act as one player and therefore, each agent can only be a member of one coalition. Over time, any number of coexisting but disjoint coalitions may emerge, each of which has to be a connected set of nodes in the network, i.e., formally a coalition structure is a partition of the network nodes into connected sets[24]. Each coalition i generates a payoff flow Π_x for each of its members x given by

$$\Pi_x s_i = (1 + X^2/2s_i)/(1 + X^2/s_i)^2 - F. \tag{2}$$

In this, s_i is the size of coalition i, F a parameter representing the fixed costs of maintaining a coalition, and X the solution to the equation

$$1 - 1/X = \sum_j 1/\left(1 + X^2/s_j\right), \tag{3}$$

where the sum is over all coalitions j (see the Supporting Information, SI, for an economic derivation of this equation from a common pool resource exploitation game). In each time-step of the model, either of two processes occur:

1. With probability ϕ, the network adapts to the coalition structure by rewiring one cross-coalitional link of a randomly chosen agent x to another randomly chosen member of x's coalition (unless x is already linked to all her coalition members). This keeps the total number of links constant which is approximately true for many real-world social systems[25].
2. Otherwise, i.e., with probability $1-\phi$, a randomly chosen agent x may change the coalition structure. The agent may either

(a). leave her coalition (in which case the rest of the coalition splits up into its connected components),
(b). merge her coalition with any combination of her neighbors' coalitions (in which case this merger must be profitable to all affected nodes in terms of the underlying payoff model),
(c). or do nothing,

depending on which of these moves results in the largest next time-step's payoff for x.

Note that the amount of change caused by one instance of process 1 is restricted to only three nodes, while process 2 typically affects a much larger number of nodes in one step, especially when the involved coalitions are already meso- or macroscopic.

The model has converged when no agents are able to rewire their links or find it profitable to change the coalition structure any longer. In the corresponding steady state there may still be several coalitions in each connected component of the network (see Fig. 2). Thus, the order parameter defining order and disorder in this socio-economic context is not the network component size but the size of the largest coalition, S. If we imagine assigning different coalitions to different spin directions, it is possible to draw an analogy to the magnetic spin model[20]. If all nodes are singletons, there are N different coalitions whose sizes does not exceed 1 (hence, $S=1$). In the analogy, all spins would be pointing into different directions averaging out to a macroscopic magnetization of zero. The other extreme would be the state of a grand coalition where $S=N$. Here, all spins would be pointing into the same direction resulting in a non-zero magnetization (the ferromagnetic state). The transition from one state of the order parameter to the other can be of first or second order. In our model, without network adaptation (for $\phi=0$) the largest coalition converges to a size S of the order of N. With increasing ϕ the coalition formation process is increasingly disturbed and S decreases. Therefore, the adaptation rate ϕ is the natural choice for the control parameter.

From the feedback loop between coalition and network structure we expect the dynamics of this model to be highly non-linear. We study these dynamics with varying control parameter ϕ, in particular the occurrence of non-equilibrium phase transitions. Phase transitions can be identified and characterized with the help of scaling theory which states characteristic system variables (order parameters) to be power-law distributed at the critical value of the control parameter[21,23]. A visualization of the coalition structure for different system sizes gives a first insight (Fig. 2). At the critical point, system patterns should not substantially change for different system sizes. As a quantification we accompany this graphical hint of finite-size scaling with the frequency distribution of all coalition sizes, s, that remain after the model has converged (Fig. 3a–c), which may take up to 10^6 time steps.

References

1. Newman, M. *Networks: An Introduction* (Oxford University Press, Inc., New York, NY, USA, 2010).
2. Gómez-Gardeñes, J., Reinares, I., Arenas, A. & Floría, L. M. Evolution of Cooperation in Multiplex Networks. *Scientific Reports* **2** (2012).
3. Gross, T. & Blasius, B. Adaptive Coevolutionary Networks: a Review. *Journal of The Royal Society Interface* **5,** 259–271 (2008).
4. Wiedermann, M., Donges, J., Heitzig, J., Lucht, W. & Kurths, J. Macroscopic Description of Complex Adaptive Networks Co-evolving with Dynamic Node States. *submitted* (2015).
5. Castellano, C., Fortunato, S. & Loreto, V. Statistical Physics of Social Dynamics. *Reviews of Modern Physics* **81,** 591–646 (2009).
6. Holme, P. & Newman, M. E. J. Nonequilibrium Phase Transition in the Coevolution of Networks and Opinions. *Phys. Rev. E* **74,** 056108 (2006).
7. Sethi, R. & Somanathan, E. The Evolution of Social Norms in Common Property Resource Use. *The American Economic Review* **86,** pp. 766–788 (1996).
8. Niwa, H.-S. School Size Statistics of Fish. *Journal of Theoretical Biology* **195,** 351–361 (1998).
9. Defran, R., Defran, W. & David, W. Occurrence, Distribution, Site Fidelity, and School Size of Bottlenose Dolphins (Tursiops Truncatus) off San Diego, California. *Marine Mammal Science* **15,** 366–380 (1999).
10. Wu, Y., Zhou, C., Xiao, J., Kurths, J. & Schellnhuber, H. J. Evidence for a Bimodal Distribution in Human Communication. *Proceedings of the National Academy of Sciences* **107,** 18803–18808 (2010).
11. Soo, K. T. Zipf's Law for Cities: a Cross-country Investigation. *Regional Science and Urban Economics* **35,** 239–263 (2005).
12. Chakrabarti, A. S. Bimodality in the Firm Size Distributions: a Kinetic Exchange Model Approach. *European Physical Journal B* **86,** 255 (2013).
13. Ramsden, J. & Kiss-Haypal, G. Company size distribution in different countries. *Physica A: Statistical Mechanics and its Applications* **277,** 220–227 (2000).
14. Family, F. & Meakin, P. Kinetics of Droplet Growth Processes: Simulations, Theory, and Experiments. *Physical Review A* **40,** 3836 (1989).

15. Stanley, H. E. *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press 1 (1987).
16. Domb, C. F. *The Critical Point: a Historical Introduction to the Modern Theory of Critical Phenomena* (Taylor & Francis London, 1996).
17. Cabral, L. M. & Mata, J. On the Evolution of the Firm Size Distribution: Facts and Theory. *American Economic Review* 1075–1090 (2003).
18. Mouritsen, O. G. *Computer Studies of Phase Transitions and Critical Phenomena*. Springer Series in Computational Physics (Springer, 1984).
19. Sornette, D. *Critical Phenomena in Natural Sciences: Chaos, Fractals, Selforganization and Disorder: Concepts and Tools* (Springer Science & Business, 2006).
20. Hinrichsen, H. Non-equilibrium Phase Transitions. *Physica A: Statistical Mechanics and its Applications* **369,** 1–28 (2006).
21. Privman, V. *Finite Size Scaling and Numerical Simulation of Statistical Systems* (World Scientific Singapore, 1990).
22. Pfeuty, P. & Toulouse, G. Introduction to the Renormalization Group and to Critical Phenomena. *Physics Today* **31,** 57 (1978).
23. Schöll, E. *Nonequilibrium Phase Transitions in Semiconductors: Self-Organization Induced by Generation and Recombination Processes. Springer Series in Synergetics* (Springer-Verlag, 1987).
24. Ray, D. & Vohra, R. A Theory of Endogenous Coalition Structures. *Games and Economic Behavior* **26,** 286–336 (1999).
25. Jin, E. M., Girvan, M. & Newman, M. E. Structure of Growing Social Networks. *Physical Review E* **64,** 046132 (2001).

Acknowledgments

This work was carried out within the framework of PIK's COPAN project. We gratefully acknowledge funding by the German Federal Ministry of Education and Research (BMBF) via the CoNDyNet project, grant no. 03SF0472A, and of the Government of the Russian Federation (Agreement No. 14.Z50.31.0033 with Institute of Applied Physics RAS). We thank Jonathan Donges for helpful discussions on this manuscript.

Author Contributions

J.H., S.A., E.S. and J.K. conceived the model and the experiments, S.A. conducted the experiments, S.A., U.K. and J.H. analyzed the results, S.A., J.H. and U.K. wrote the manuscript, all authors discussed the results and designed and reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Auer, S. *et al.* The Dynamics of Coalition Formation on Complex Networks. *Sci. Rep.* **5**, 13386; doi: 10.1038/srep13386 (2015).

Reproduced from: S. Auer et al., The Dynamics of Coalition Formation on Complex Networks, Sci. Rep., vol. 5, pp. 1–8, 2015, doi: 10.1038/srep13386. Published under Creative Commons Attribution License 4.0 (CC BY).

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Erratum: The Dynamics of Coalition Formation on Complex Networks

S. Auer, J. Heitzig, U. Kornek, E. Schöll & J. Kurths

Correction to: *Scientific Reports* https://doi.org/10.1038/srep13386; published online 25 August 2015; updated 11 May 2018

In the original version of this Article, the Supplementary Information file containing the derivation for individual payoffs was omitted. This error has now been corrected in the PDF and HTML versions of the Article.

PHYSICAL REVIEW E **91**, 052801 (2015)

Macroscopic description of complex adaptive networks coevolving with dynamic node states

Marc Wiedermann,[1,2,*] Jonathan F. Donges,[1,3] Jobst Heitzig,[1] Wolfgang Lucht,[1,4] and Jürgen Kurths[1,2,5,6]

[1] *Potsdam Institute for Climate Impact Research, P. O. Box 60 12 03, 14412 Potsdam, Germany, EU*
[2] *Department of Physics, Humboldt University, Newtonstr. 15, 12489 Berlin, Germany, EU*
[3] *Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden, EU*
[4] *Department of Geography, Humboldt University, Rudower Chaussee 16, 12489 Berlin, Germany, EU*
[5] *Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3FX, United Kingdom, EU*
[6] *Department of Control Theory, Nizhny Novgorod State University, Gagarin Avenue 23, 606950 Nizhny Novgorod, Russia*

(Received 16 December 2014; published 7 May 2015)

In many real-world complex systems, the time evolution of the network's structure and the dynamic state of its nodes are closely entangled. Here we study opinion formation and imitation on an adaptive complex network which is dependent on the individual dynamic state of each node and vice versa to model the coevolution of renewable resources with the dynamics of harvesting agents on a social network. The adaptive voter model is coupled to a set of identical logistic growth models and we mainly find that, in such systems, the rate of interactions between nodes as well as the adaptive rewiring probability are crucial parameters for controlling the sustainability of the system's equilibrium state. We derive a macroscopic description of the system in terms of ordinary differential equations which provides a general framework to model and quantify the influence of single node dynamics on the macroscopic state of the network. The thus obtained framework is applicable to many fields of study, such as epidemic spreading, opinion formation, or socioecological modeling.

DOI: 10.1103/PhysRevE.91.052801 PACS number(s): 89.75.Hc, 87.23.Ge, 89.65.−s, 89.75.Fb

I. INTRODUCTION

Complex network theory has proven to be a powerful tool for studying properties, dynamics, and evolution of many real-world complex systems [1,2]. Of particular interest is the ability to investigate adaptive or temporal networks and their respective dynamics [3–5]. Typical processes studied in this field are epidemic spreading [6–8] or opinion formation, e.g., based on the adaptive voter model [9,10]. Interactions are modeled by randomly picking a pair of linked nodes and, with fixed probabilities, either changing the state of one of the two nodes or modifying their neighborhood structure by adaptive rewiring. However, recent results have emphasized that opinion formation and imitation processes in fact do not take place with fixed probabilities but can depend on the payoff or performance of different opinion-related choices made by the agents or nodes involved [11–13].

In addition to the structure and dynamics *of* networks there have been a variety of studies on the dynamics *on* networks, where nodes in the network represent individual dynamical systems and links indicate directed or symmetric interactions between them [14,15]. It has been suggested that the interplay between the dynamics of and on networks should be much more thoroughly investigated, since the dynamics of each of the coupled subsystems is expected to change significantly when compared to their autonomous time evolution [3].

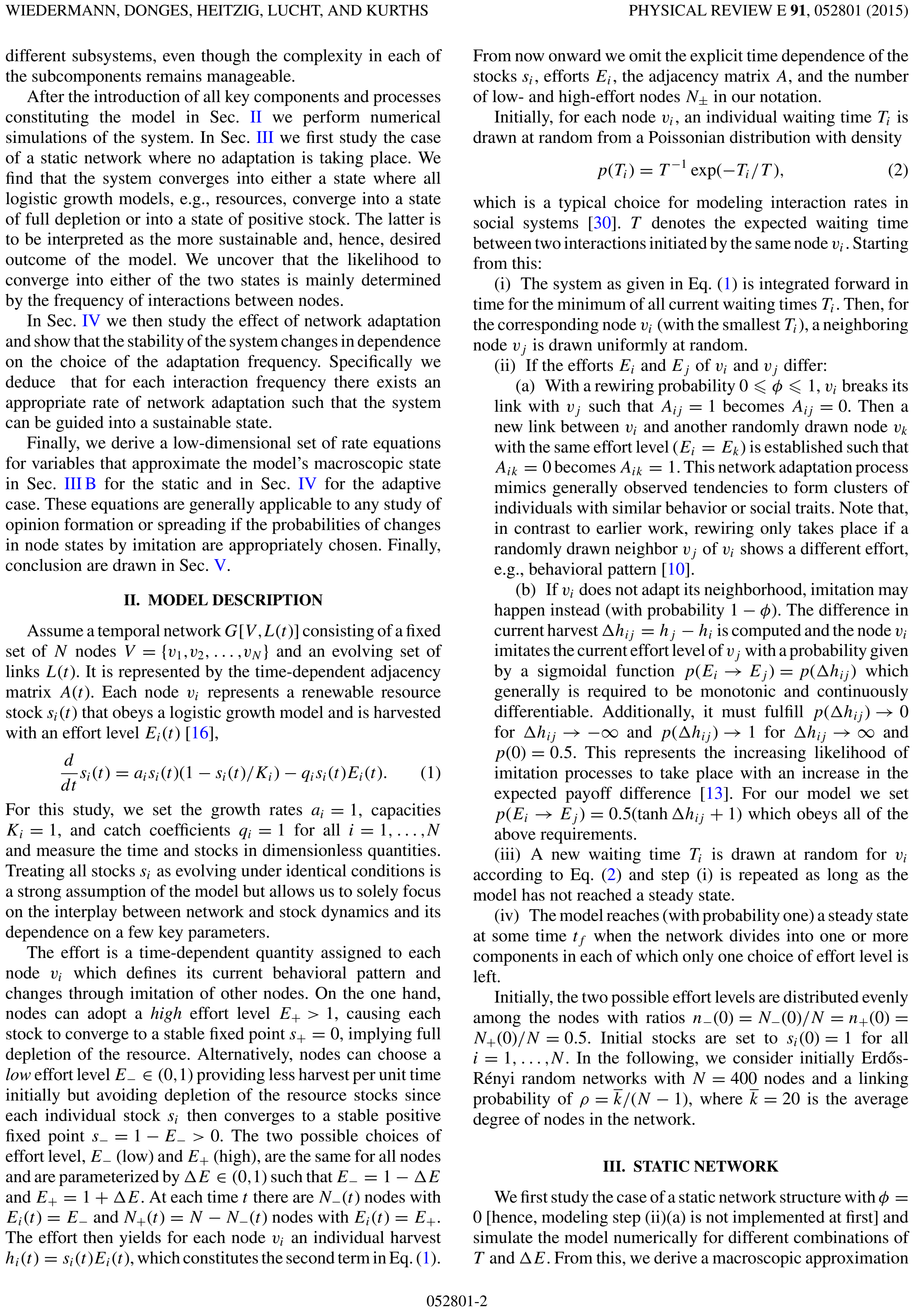
In this work, we propose a model that combines both aspects. For this purpose we refine the adaptive voter model so there is no fixed probability for pairs of nodes to either imitate each other's opinion or adaptively rewire their acquaintance structure. Instead, each node also represents a dynamical system which, for illustration, is chosen here to be simple and easily understood if treated in an isolated fashion. In particular, we choose a logistic growth model, which is a paradigm for the dynamics of a bounded renewable resource [16]. Whenever interactions between nodes take place, the states of the respective dynamical systems are also taken into account. As a consequence, imitation processes depend explicitly on the nodes' states as well as on the current network structure. At the same time each of the nodes' opinions influences a parameter of the local dynamical system.

The proposed model serves as a narrative for possibly emerging dynamics in coevolutionary human-nature interactions [17–19]. It complements conceptual studies on the effects of economic growth on the ecospheric state [20,21] as well as work on resource exploitation models that take into account the coevolution of stylized resource dynamics with a similarly paradigmatic population growth model [22,23]. The proposed model, for the first time, takes into account individual pairwise interactions of agents on a social network when studying the stability and dynamics of such intertwined systems.

So far, in the context of sustainability science [24], studies on the effect of different exploitation strategies on the state of a certain ecospheric component have been carried out by, e.g., studying the extraction of water in rivers by a network of interconnected harvesters [25–27]. However, no systematic analysis of the underlying network structure and resulting dynamics was performed. In addition, no network dynamics, such as adaptation or imitation processes, were included in these studies and the focus was mainly set on studying the state of the ecosphere for different harvesting strategies that were evolving deterministically in order to optimize all harvesters' payoffs.

In contrast, imitation dynamics with high numbers of agents or players have been studied in the context of evolutionary game theory [12,13,28,29]. However, in no such cases were the dynamics of resources or other externalities taken into account and, hence, no coevolution of different subsystems has been studied. Here, the proposed model serves to illustrate the rich dynamics that may emerge from the coupling of these

[*] marcwie@pik-potsdam.de



different subsystems, even though the complexity in each of the subcomponents remains manageable.

After the introduction of all key components and processes constituting the model in Sec. II we perform numerical simulations of the system. In Sec. III we first study the case of a static network where no adaptation is taking place. We find that the system converges into either a state where all logistic growth models, e.g., resources, converge into a state of full depletion or into a state of positive stock. The latter is to be interpreted as the more sustainable and, hence, desired outcome of the model. We uncover that the likelihood to converge into either of the two states is mainly determined by the frequency of interactions between nodes.

In Sec. IV we then study the effect of network adaptation and show that the stability of the system changes in dependence on the choice of the adaptation frequency. Specifically we deduce that for each interaction frequency there exists an appropriate rate of network adaptation such that the system can be guided into a sustainable state.

Finally, we derive a low-dimensional set of rate equations for variables that approximate the model's macroscopic state in Sec. III B for the static and in Sec. IV for the adaptive case. These equations are generally applicable to any study of opinion formation or spreading if the probabilities of changes in node states by imitation are appropriately chosen. Finally, conclusion are drawn in Sec. V.

II. MODEL DESCRIPTION

Assume a temporal network $G[V,L(t)]$ consisting of a fixed set of N nodes $V = \{v_1, v_2, \dots, v_N\}$ and an evolving set of links $L(t)$. It is represented by the time-dependent adjacency matrix $A(t)$. Each node v_i represents a renewable resource stock $s_i(t)$ that obeys a logistic growth model and is harvested with an effort level $E_i(t)$ [16],

$$\frac{d}{dt}s_i(t) = a_i s_i(t)(1 - s_i(t)/K_i) - q_i s_i(t) E_i(t). \qquad (1)$$

For this study, we set the growth rates $a_i = 1$, capacities $K_i = 1$, and catch coefficients $q_i = 1$ for all $i = 1, \dots, N$ and measure the time and stocks in dimensionless quantities. Treating all stocks s_i as evolving under identical conditions is a strong assumption of the model but allows us to solely focus on the interplay between network and stock dynamics and its dependence on a few key parameters.

The effort is a time-dependent quantity assigned to each node v_i which defines its current behavioral pattern and changes through imitation of other nodes. On the one hand, nodes can adopt a *high* effort level $E_+ > 1$, causing each stock to converge to a stable fixed point $s_+ = 0$, implying full depletion of the resource. Alternatively, nodes can choose a *low* effort level $E_- \in (0,1)$ providing less harvest per unit time initially but avoiding depletion of the resource stocks since each individual stock s_i then converges to a stable positive fixed point $s_- = 1 - E_- > 0$. The two possible choices of effort level, E_- (low) and E_+ (high), are the same for all nodes and are parameterized by $\Delta E \in (0,1)$ such that $E_- = 1 - \Delta E$ and $E_+ = 1 + \Delta E$. At each time t there are $N_-(t)$ nodes with $E_i(t) = E_-$ and $N_+(t) = N - N_-(t)$ nodes with $E_i(t) = E_+$. The effort then yields for each node v_i an individual harvest $h_i(t) = s_i(t)E_i(t)$, which constitutes the second term in Eq. (1). From now onward we omit the explicit time dependence of the stocks s_i, efforts E_i, the adjacency matrix A, and the number of low- and high-effort nodes $N_\pm$ in our notation.

Initially, for each node v_i, an individual waiting time T_i is drawn at random from a Poissonian distribution with density

$$p(T_i) = T^{-1}\exp(-T_i/T), \qquad (2)$$

which is a typical choice for modeling interaction rates in social systems [30]. T denotes the expected waiting time between two interactions initiated by the same node v_i. Starting from this:

(i) The system as given in Eq. (1) is integrated forward in time for the minimum of all current waiting times T_i. Then, for the corresponding node v_i (with the smallest T_i), a neighboring node v_j is drawn uniformly at random.

(ii) If the efforts E_i and E_j of v_i and v_j differ:

(a) With a rewiring probability $0 \leqslant \phi \leqslant 1$, v_i breaks its link with v_j such that $A_{ij} = 1$ becomes $A_{ij} = 0$. Then a new link between v_i and another randomly drawn node v_k with the same effort level ($E_i = E_k$) is established such that $A_{ik} = 0$ becomes $A_{ik} = 1$. This network adaptation process mimics generally observed tendencies to form clusters of individuals with similar behavior or social traits. Note that, in contrast to earlier work, rewiring only takes place if a randomly drawn neighbor v_j of v_i shows a different effort, e.g., behavioral pattern [10].

(b) If v_i does not adapt its neighborhood, imitation may happen instead (with probability $1 - \phi$). The difference in current harvest $\Delta h_{ij} = h_j - h_i$ is computed and the node v_i imitates the current effort level of v_j with a probability given by a sigmoidal function $p(E_i \to E_j) = p(\Delta h_{ij})$ which generally is required to be monotonic and continuously differentiable. Additionally, it must fulfill $p(\Delta h_{ij}) \to 0$ for $\Delta h_{ij} \to -\infty$ and $p(\Delta h_{ij}) \to 1$ for $\Delta h_{ij} \to \infty$ and $p(0) = 0.5$. This represents the increasing likelihood of imitation processes to take place with an increase in the expected payoff difference [13]. For our model we set $p(E_i \to E_j) = 0.5(\tanh \Delta h_{ij} + 1)$ which obeys all of the above requirements.

(iii) A new waiting time T_i is drawn at random for v_i according to Eq. (2) and step (i) is repeated as long as the model has not reached a steady state.

(iv) The model reaches (with probability one) a steady state at some time t_f when the network divides into one or more components in each of which only one choice of effort level is left.

Initially, the two possible effort levels are distributed evenly among the nodes with ratios $n_-(0) = N_-(0)/N = n_+(0) = N_+(0)/N = 0.5$. Initial stocks are set to $s_i(0) = 1$ for all $i = 1, \dots, N$. In the following, we consider initially Erdős-Rényi random networks with $N = 400$ nodes and a linking probability of $\rho = \bar{k}/(N-1)$, where $\bar{k} = 20$ is the average degree of nodes in the network.

III. STATIC NETWORK

We first study the case of a static network structure with $\phi = 0$ [hence, modeling step (ii)(a) is not implemented at first] and simulate the model numerically for different combinations of T and ΔE. From this, we derive a macroscopic approximation

Reprinted paper with permission from M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 91, no. 5, pp. 1-11, 2015, doi: 10.1103/PhysRevE.91.052801. Copyright 2015 by the American Physical Society.

of the model constituted from a set of three coupled differential equations and show its good agreement with the numerical results.

A. Numerical simulations

Numerical simulations for different combinations of T and ΔE provide insights into the system's dynamics. Figure 1(a) shows the fraction $f_-(t_f)$ of model runs that converge to a state where all nodes show a low effort $E_i(t_f) = E_- \; \forall\, i = 1,\ldots,N$ (using an ensemble of $n = 500$ simulations). For small T (fast interactions) there is a high probability for the system to converge to a state where only nodes with a high effort level E_+ are present. In this case all resource stocks converge to the stable fixed point $s_+ = 0$ and become fully depleted. With increasing T, the system's expected equilibrium state undergoes a phase transition in $f_-(t_f)$. For sufficiently large T (slow interactions), the system is likely to converge to a state where all nodes adopt the effort level E_- and all stocks converge to a stable fixed point $s_- = 1 - E_- > 0$. This indicates that the rate of interactions between nodes plays a crucial role in determining the system's expected equilibrium state.

The resulting dynamics can be qualitatively understood by considering the limiting cases of $T \to 0$ and $T \to \infty$. In the first case, interactions between nodes are expected to happen very fast. Given that initially all stocks carry the same value $s_i(0) = s_0$ we expect that for $t \ll 1$ the harvest h_- (h_+) of nodes with low (high) effort follows $h_-(t \ll 1) \propto E_- s_0$ [$h_+(t \ll 1) \propto E_+ s_0$]. This implies that the difference in harvest between the two different types of nodes is expected as $h_+(t \ll 1) - h_-(t \ll 1) \propto (E_+ - E_-)s_0 = 2\Delta E s_0$. If interactions happen very fast, the system likely converges into its equilibrium state at $t_f \ll 1$. Since in this situation we expect $h_+ > h_-$, nodes with low effort are more likely to imitate the high effort rather than the other way around and, hence, we expect $f_-(t_f) \to 0$ for $T \to 0$ [as can be seen in Fig. 1(a)].

In contrast, for $T \to \infty$, we expect updates between nodes to happen preferably at times $t \gg 1$. In this case, the stocks of nodes with high (low) effort can be assumed to have already converged to a fixed point of $s_+ = 0$ ($s_- = 1 - E_+ = \Delta E$) as interactions between nodes start to take place. Hence, the difference in harvest is expected as $h_-(t \gg 1) - h_+(t \gg 1) = \Delta E - \Delta E^2$. Thus, for all $\Delta E \in (0,1)$ the harvest of low-effort nodes exceeds that of nodes with high effort and the system is likely to converge into a state where all nodes show the low effort and, hence, $f_-(t_f) \to 1$ [red (dark) area in Fig. 1(a) for high values of T].

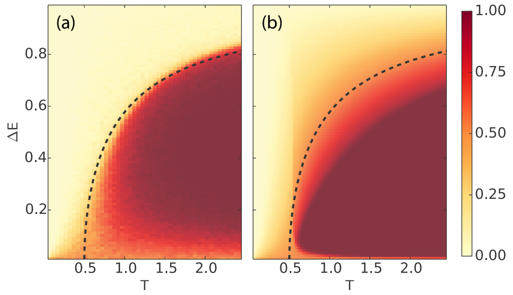


FIG. 1. (Color online) (a) The fraction $f_-(t_f)$ of numerical simulations that converge to a state where all nodes show a low effort level $E_i(t_f) = E_- \; \forall\, i = 1,\ldots,N$ computed over $n = 500$ runs for different choices of T and ΔE for a static network with $\phi = 0$. (b) The value n_{-0} of the stable fixed point for the fraction n_- of nodes with effort level E_- computed from Eqs. (16)–(18). The dashed line indicates the critical expected waiting time T_c which separates the two regimes (predominance of nodes using E_+ [yellow (light)] and E_- [red (dark)].

We note that $h_-(t \gg 1) - h_+(t \gg 1) = \Delta E - \Delta E^2$ varies with ΔE. Specifically, in the limiting cases $\Delta E = 0$ and $\Delta E = 1$ we find that the difference $h_-(t \gg 1) - h_+(t \gg 1) = 0$ vanishes and, hence, the system becomes equally likely to converge into either a state with only low-effort nodes or only high-effort nodes present [see lower right corner and the shift of the transition point towards higher T with increasing ΔE in Fig. 1(a)].

B. Macroscopic approximation

Abstracting from pairwise microscopic interactions, we now look at the system from a macroscopic point of view. Assuming the network to be large and fully connected at first, the time evolution of the system's state can be characterized by rate equations for three quantities: (1) the fraction of nodes n_- with effort level E_-, (2) the mean resource stock $\mu_- = \langle s_i | E_i = E_- \rangle_i$ of nodes with effort level E_-, and (3) the mean resource stock $\mu_+ = \langle s_i | E_i = E_+ \rangle_i$ of nodes with effort level E_+. The fraction of nodes n_+ with effort level E_+ follows from $n_+ = 1 - n_-$.

The time evolution of n_- is governed by nodes that change from the low to the high effort level and vice versa. In particular, in the time interval $(t, t + dt)$ an infinitesimal fraction of $dn_{-\to+}$ ($dn_{+\to-}$) nodes change their effort from E_- (E_+) to E_+ (E_-), which decreases (increases) the fraction of nodes with low effort n_-,

$$dn_- = dn_{+\to-} - dn_{-\to+}. \tag{3}$$

The interactions between nodes that govern the rates of changes in effort are driven by the following quantities:

(1) The expected waiting time T for a node v_i to interact with a randomly drawn neighboring node v_j. Correspondingly, the rate of node interactions is taken to be $\tau = 1/T$.

(2) If a node v_i interacts with its neighboring node v_j, an imitation of effort only takes place if $E_i \neq E_j$. Hence, for a node v_i with $E_i = E_-$ ($E_i = E_+$) there is to define a probability P_-^+ (P_+^-) that a randomly drawn neighboring node v_j has $E_j = E_+$ ($E_j = E_-$). Since a large fully connected network is assumed, this probability is given exactly by the current fraction n_+ (n_-) of nodes with high (low) effort E_+ (E_-) and, hence, $P_-^+ = n_+$ ($P_+^- = n_-$).

(3) If a node v_i with $E_i = E_-$ ($E_i = E_+$) interacts with a neighboring node v_j with $E_j = E_+$ ($E_j = E_-$), there is a probability $p_{-\to+}$ ($p_{+\to-}$) that v_i takes up the effort level E_j of v_j. This probability is governed by the difference in harvest Δh_{ij} between v_j and v_i. For the macroscopic description, the individual pairwise interactions are replaced by aggregated quantities. Therefore $p_{-\to+}$ ($p_{+\to-}$) is computed as the *expected* probability for a node v_i with low (high) effort to adopt the high (low) effort given that it interacts with a node v_j that currently has $E_j = E_+$ ($E_j = E_-$). This quantity is then dependent on the expected stocks at nodes with low and high effort, which is derived below in detail.

Reprinted paper with permission from M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 91, no. 5, pp. 1-11, 2015, doi: 10.1103/PhysRevE.91.052801. Copyright 2015 by the American Physical Society.

This yields $dn_{-\to+}$ and $dn_{+\to-}$ as the product of all three factors introduced above,

$$dn_{-\to+} = n_-\tau n_+ p_{-\to+} dt, \tag{4}$$

$$dn_{+\to-} = n_+\tau n_- p_{+\to-} dt, \tag{5}$$

$$\Rightarrow \frac{dn_-}{dt} = \tau n_- n_+ (p_{+\to-} - p_{-\to+}). \tag{6}$$

The two quantities still remaining to be evaluated are the expected probabilities $p_{+\to-}$ $(p_{-\to+})$ for nodes with a high (low) effort level to change to the opposite level. It is obtained as the expected probability for nodes in the network to take up its neighbor's effort,

$$\begin{aligned}
p_{+\to-} &= \langle p(E_j \to E_k) | E_j = E_+,\, E_k = E_- \rangle_{j,k} \\
&= 0.5 \langle \tanh(\Delta h_{jk} | E_j = E_+,\, E_k = E_-) \rangle_{j,k} + 0.5 \\
&\cong 0.5 \langle \Delta h_{jk} | E_j = E_+,\, E_k = E_- \rangle_{j,k} + 0.5 \\
&= 0.5 (E_- \langle s_k | E_k = E_- \rangle_k - E_+ \langle s_j | E_j = E_+ \rangle_j) + 0.5 \\
&= 0.5 (E_- \mu_- - E_+ \mu_+) + 0.5 \qquad (7)\\
p_{-\to+} &= 0.5 (E_+ \mu_+ - E_- \mu_-) + 0.5. \qquad (8)
\end{aligned}$$

Here we performed a linear expansion of the hyperbolic tangent, $\tanh x = x + O(x^3)$, assuming that differences in harvest remain small.

The time evolution of either of the two average stocks μ_- and μ_+ is governed by two terms. First, each individual stock s_i follows the logistic growth model and so do the average quantities. Second, the value of each of the two average stocks changes according to the fact that the nodes modify their effort from E_- to E_+ and vice versa during the time interval $(t, t + dt)$. This yields

$$\begin{aligned}
d\mu_- &= d\langle s_k | E_k = E_- \rangle_k \\
&= \langle ds_k | E_k = E_- \rangle_k \\
&= dt\ \langle s_k(1 - s_k) - E_k s_k | E_k = E_- \rangle_k + \delta_- \\
&= dt\mu_- - dt \langle s_k^2 | E_k = E_- \rangle_k - dt E_- \mu_- + \delta_- \\
&= dt(\mu_-(1 - \mu_- - E_-) - \mu_-^{(2)}) + \delta_- \qquad (9)\\
d\mu_+ &= dt(\mu_+(1 - \mu_+ - E_+) - \mu_+^{(2)}) + \delta_+. \qquad (10)
\end{aligned}$$

Here $\mu_-^{(2)}$ and $\mu_+^{(2)}$ denote the variances in the two types of stocks. δ_- (δ_+) indicate the net change in the average stock as nodes with high (low) effort change their effort to the opposite choice during $(t, t + dt)$. The fraction of nodes $dn_{+\to-}$ $(dn_{-\to+})$ that change their effort from E_+ to E_- (E_- to E_+) during $(t, t + dt)$ is assumed to be small compared to the fraction of nodes which already hold the low (high) effort, $dn_{+\to-} \ll n_-$ $(dn_{-\to+} \ll n_+)$. Hence, the respective contribution to the dynamics of μ_- (μ_+) as nodes change their effort is also assumed to be small, $dn_{+\to-}\mu_+ \ll n_-\mu_-$ $(dn_{-\to+}\mu_- \ll n_+\mu_+)$. This allows for a first-order expansion of the stock's time evolution, such that

$$\begin{aligned}
\mu_- + \delta_- &= \frac{(n_- - dn_{-\to+})\mu_- + dn_{+\to-}\mu_+}{n_- - dn_{-\to+} + dn_{+\to-}} \\
&\cong \left. \frac{(n_- - dn_{-\to+})\mu_- + dn_{+\to-}\mu_+}{n_- - dn_{-\to+} + dn_{+\to-}} \right|_{(dn_{-\to+}, dn_{+\to-}) = (0,0)} \\
&\quad + \left. \frac{-\mu_-(n_- - dn_{-\to+} + dn_{+\to-}) + ((n_- - dn_{-\to+})\mu_- + dn_{+\to-}\mu_+)}{(n_- - dn_{-\to+} + dn_{+\to-})^2} \right|_{(dn_{-\to+}, dn_{+\to-}) = (0,0)} dn_{-\to+} \\
&\quad + \left. \frac{\mu_+(n_- - dn_{-\to+} + dn_{+\to-}) - ((n_- - dn_{-\to+})\mu_- + dn_{+\to-}\mu_+)}{(n_- - dn_{-\to+} + dn_{+\to-})^2} \right|_{(dn_{-\to+}, dn_{+\to-}) = (0,0)} dn_{+\to-} \\
&= \frac{n_-\mu_-}{n_-} + \frac{-\mu_- n_- + n_-\mu_-}{n_-^2} dn_{-\to+} + \frac{\mu_+ n_- - n_-\mu_-}{n_-^2} dn_{+\to-} = \mu_- + \frac{\mu_+ - \mu_-}{n_-} dn_{+\to-} \qquad (11)\\
\Rightarrow \delta_- &= (\mu_+ - \mu_-) n_+ \tau p_{+\to-} dt \qquad (12)\\
\delta_+ &= (\mu_- - \mu_+) n_- \tau p_{-\to+} dt. \qquad (13)
\end{aligned}$$

Putting this back into (9) and (10) yields

$$d\mu_- = dt(\mu_-(1 - \mu_- - E_-) - \mu_-^{(2)}) + dt(\mu_+ - \mu_-) n_+ \tau p_{+\to-} \tag{14}$$

$$d\mu_+ = dt(\mu_+(1 - \mu_+ - E_+) - \mu_+^{(2)}) + dt(\mu_- - \mu_+) n_- \tau p_{-\to+}. \tag{15}$$

In the scope of this work, in to order to close the set of equations that describe the systems dynamics, we assume the respective variances $\mu_-^{(2)}$ and $\mu_+^{(2)}$ to vanish. Taking into account higher moments in the dynamics of the stocks and investigate its influence on the resulting fixed points remains as a task for future research.

In summary, we find a set of three coupled ordinary differential equations that define the time evolution of the static network model:

$$\frac{dn_-}{dt} = \tau n_+ n_- (p_{+\to-} - p_{-\to+}) \tag{16}$$

$$\frac{d\mu_-}{dt} = \mu_-(1-\mu_- - E_-) + \tau(\mu_+ - \mu_-)n_+ p_{+\to -} \tag{17}$$

$$\frac{d\mu_+}{dt} = \mu_+(1-\mu_+ - E_+) + \tau(\mu_- - \mu_+)n_- p_{-\to +}. \tag{18}$$

C. Fixed points and stability

We obtain all fixed points $P_i = (n_{-0}, \mu_{-0}, \mu_{+0})$ of the dynamical system given in Eqs. (16)–(18) as:

$$P_1 = \left(n_{-0} = 0,\ \mu_{-0} = \frac{1 - E_- - 0.5\tau}{1 + 0.5\tau E_-},\ \mu_{+0} = 0\right), \tag{19}$$

$$P_2 = \left(n_{-0} = 1,\ \mu_{-0} = 0,\ \mu_{+0} = \frac{1 - E_+ - 0.5\tau}{1 + 0.5\tau E_+}\right), \tag{20}$$

$$P_3 = \left[n_{-0} = \frac{2\left(E_- \frac{1-0.5\tau}{E_- + E_+} + E_+ - 1\right)}{\tau\left(\frac{E_+}{E_-} - 1\right)},\ \mu_{-0} = E_+ \frac{1 - 0.5\tau}{E_- + E_+},\ \mu_{+0} = E_- \frac{1 - 0.5\tau}{E_- + E_+}\right], \tag{21}$$

$$P_4 = \left[n_{-0} = 1,\ \mu_{-0} = 1 - E_-,\ \mu_{+0} = \frac{-b}{2a} + \sqrt{\left(\frac{b}{2a}\right)^2 + \frac{c}{a}}\right], \tag{22}$$

$$P_5 = \left[n_{-0} = 1,\ \mu_{-0} = 1 - E_-,\ \mu_{+0} = \frac{-b}{2a} - \sqrt{\left(\frac{b}{2a}\right)^2 + \frac{c}{a}}\right], \tag{23}$$

$$\begin{aligned} a &= 0.5(-2 - E_+\tau) \\ b &= 1 - E_+ + 0.5\tau[(1 - E_-)E_+ + E_- - E_-^2 - 1] \\ c &= 0.5\tau(1 - E_-)(E_- - E_-^2 - 1). \end{aligned}$$

In addition, there exists a manifold which also satisfies $\frac{dn_-}{dt} = \frac{d\mu_-}{dt} = \frac{d\mu_+}{dt} = 0$ and is defined by

$$P_\alpha = (n_{-0} = \alpha,\ \mu_{-0} = 0,\ \mu_{+0} = 0),\ \alpha \in [0,1]. \tag{24}$$

For all fixed points given above we compute the largest eigenvalue λ_1 of the corresponding Jacobian matrix evaluated at the respective point. Only the two fixed points P_3 and P_4 have a negative largest eigenvalue $\lambda_1 < 0$ and, hence, are stable for choices of parameters ΔE and $T > 0.5$ (note that, again, $E_- = 1 - \Delta E$, $E_+ = 1 + \Delta E$, and $\tau = 1/T$) (Fig. 2).

To investigate the system's dynamics in the regime $T < 0.5$, the stability on the one-dimensional manifold defined by all points that fulfill Eq. (24) is assessed. Analytically computing the three eigenvalues of the Jacobian matrix on the manifold as a function of the parameter α yields

$$\lambda_0 = 0, \tag{25}$$

$$\lambda_\pm(\alpha) = 1 - \frac{E_+ + E_-}{2} - \frac{\tau}{4} \pm \frac{1}{2}\sqrt{2\alpha E_+\tau - 2\alpha E_-\tau + E_+^2 - 2E_+E_- - E_+\tau + E_-^2 + E_-\tau + \frac{\tau^2}{4}}. \tag{26}$$

A first observation is that $\lambda_+(\alpha) \geqslant \lambda_-(\alpha)$ holds. Since $\lambda_0 = 0$, it is obvious that not all eigenvalues can be negative. However, if $\lambda_0 = 0$ is the largest eigenvalue of the system, all choices of α for which $\lambda_+(\alpha) \leqslant \lambda_0$ define a center manifold,

$$\lambda_+(\alpha) \leqslant 0 \quad \text{if} \quad \alpha \leqslant \tfrac{1}{2} - T\Delta E. \tag{27}$$

Hence,

$$\begin{aligned} \nu(\alpha) &= (n_{-0} = \alpha, \mu_{-0} = 0, \mu_{+0} = 0) \\ \alpha &\in \left[0, \tfrac{1}{2} - T\Delta E\right] \end{aligned} \tag{28}$$

defines a center manifold where the system's stability cannot be assessed by linear stability analysis. A detailed study of the system's stability in this regime is beyond the scope of this work and not necessarily needed to understand the general dynamics of the macroscopic description proposed here. Numerically integrating the system for choices of parameters taken from the center manifold, however, reveals good agreement between the microscopic and macroscopic model representation (Fig. 1). An investigation by means of a higher-order stability analysis might yield further insights into the processes that cause both resource stocks $\mu_{-0} = \mu_{+0} = 0$ to be fully depleted in the regime of the center manifold.

Reprinted paper with permission from M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 91, no. 5, pp. 1-11, 2015, doi: 10.1103/PhysRevE.91.052801. Copyright 2015 by the American Physical Society.

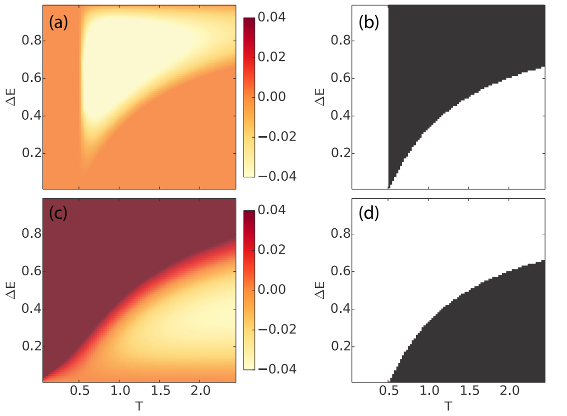


FIG. 2. (Color online) The largest eigenvalue λ_1 for the two fixed points P_3 (a) and P_4 (c) [see also Eqs. (21) and (22)] depending on ΔE and T. The black area in (b) indicates the domain in parameter space for which λ_1 computed for P_3 is negative and, hence, P_3 is stable. (d) shows the same properties for P_4. The regimes for which either of the two fixed points is stable are complementary. Further it should be noted that for $T < 0.5$ neither of the two fixed points is stable, but the center manifold as given in Eq. (28) exists in this regime.

In conclusion, we note that for each choice of parameters only one of the fixed points P_3 and P_4 can be the unique stable fixed point of the system (Fig. 2). Figure 1(b) displays the value of the stable fixed point's n_{-0} component as a function of T and ΔE. The results are in good agreement with the numerical findings [Fig. 1(a)]. Due to the first-order approximation, the transition from a predominance of nodes with E_+ to nodes with E_- with increasing T is not as sharp as for the numerical simulations. However, a good estimate for the critical value T_c of T at which the transition takes place can be found by setting $n_{-0}(T_c) = 0.5$ in Eq. (21) which yields $T_c(\Delta E) = \frac{1+\Delta E^2}{2-2\Delta E^2}$ (dashed line in Fig. 1).

IV. ADAPTIVE NETWORK

In the following, we consider additionally network adaptation processes with $\phi > 0$ [hence, modeling steps (ii)(a) and (ii)(b) both take place with a relative frequency depending on the rewiring probability ϕ]. For all results presented from here onward, the two available choices of effort levels are fixed by setting $\Delta E = 0.5$.

A. Numerical simulations

Numerical simulations with the same initial conditions as in the static case for different combinations of ϕ and T reveal a division of the parameter space into regimes of different expected outcomes as the model reaches its steady state [Fig. 3(a)]. In contrast to the static case nodes no longer necessarily all carry the same effort as the model reaches its equilibrium state, due to the possibility for the network to fragment into smaller components. Hence, from now on $f_-(t_f)$ denotes the mean fraction of nodes with effort level E_- as the model reaches consensus. As for $\phi = 0$, fast interactions (i.e., low values of T) lead to a large fraction of nodes carrying E_+. The transition between the two behavioral patterns with increasing T remains sharp. However, depending on the choice of ϕ, the value of the critical waiting time T_c, at which the system transfers from a state with a predominance of nodes with low effort to a state with a predominance of nodes with high effort decreases with increasing ϕ. Conversely, this implies that for all $T \gtrsim 0.3$ there is an appropriate choice of $\phi \in [\phi_{c_1}, \phi_{c_2}]$ so that all nodes are likely to adopt the effort level E_-. In the limiting case of $\phi = 1$ the expected fraction of nodes with E_- equals the initial fraction $n_-(0) = 0.5$ for all choices of T due to the network's fragmentation into components of nodes sharing the same effort.

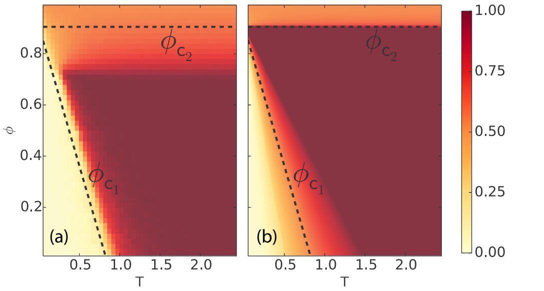


FIG. 3. (Color online) (a) Mean fraction of nodes $f_-(t_f)$ with effort level $E_- = 1 - \Delta E = 0.5$ for different choices of T and ϕ obtained from an ensemble of $n = 500$ simulations as the system reaches its steady state. (b) Value of the stable fixed point for the fraction of nodes with effort level E_- computed from the set of differential equations (59)–(63).

B. Macroscopic approximation

The macroscopic approximations (16)–(18) can be extended to also include the effects of network rewiring. For this, we introduce two additional variables describing the macroscopic state of the network. The time evolution of the fraction of nodes n_- with low effort is recalled [analogously to Eq. (6)] as

$$\frac{dn_-}{dt} = \tau(n_+ P_+^- p_{+\to-} - n_- P_-^+ p_{-\to+}). \tag{29}$$

Given that a node v_i initializes an interaction and the randomly drawn neighboring node v_j employs a different effort, $E_i \neq E_j$, there exists the adaptive rewiring probability $\phi \in [0,1]$ for v_i to break its connection with v_j and establish a link with another randomly drawn node v_k in the network that is employing the same effort as v_i ($E_k = E_i$) and is not yet connected to node v_i. With probability $1 - \phi$, imitation of efforts takes place which has already been implemented in the macroscopic description of the static network. To account for the adaptive rewiring process, the interaction rate τ needs to be refined such that it no longer represents the rate of node interactions alone, but the rate of interactions which lead to imitation,

$$\tau = \frac{1-\phi}{T}. \tag{30}$$

Reprinted paper with permission from M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 91, no. 5, pp. 1-11, 2015, doi: 10.1103/PhysRevE.91.052801. Copyright 2015 by the American Physical Society.

Likewise the ratio ρ of all node interactions that lead to adaptive rewiring needs to be defined. Since each node is expected to interact at a rate $1/T$ it follows that

$$\rho = \frac{\phi}{T}. \tag{31}$$

For adaptive rewiring to take place, the network cannot be fully connected. Therefore, the previous definitions of $P_+^- = n_-$ and $P_-^+ = n_+$ for two nodes of different effort to interact no longer hold for the derivations to be performed here.

The total number of M links in the network splits into M_- (M_+) links connecting two nodes with low (high) effort and M_{+-} links connecting two nodes of different efforts, such that

$$M = \frac{N\bar{k}}{2} = M_- + M_+ + M_{+-} \tag{32}$$

$$\Rightarrow \frac{dM}{dt} = \frac{dM_-}{dt} + \frac{dM_+}{dt} + \frac{dM_{+-}}{dt} = 0. \tag{33}$$

Additionally, let

$$K_-^- = \frac{2M_-}{N_-} \tag{34}$$

denote for nodes with low effort the average number of neighbors with the same effort. Likewise,

$$K_-^+ = \frac{M_{+-}}{N_-} \tag{35}$$

represents for nodes with low effort the average number of neighbors with high effort. These two quantities constitute the average degree of nodes with low effort as

$$K_- = K_-^- + K_-^+ = \frac{M_{+-} + 2M_-}{N_-}. \tag{36}$$

Likewise, the average degree K_+ of nodes with high effort is obtained from

$$K_+^+ = \frac{2M_+}{N_+}, \tag{37}$$

$$K_+^- = \frac{M_{+-}}{N_+}, \tag{38}$$

$$K_+ = \frac{M_{+-} + 2M_+}{N_+}. \tag{39}$$

For a node v_i currently having a low effort $E_i = E_-$ the probability $P_-^+(v_i)$ to draw a neighbor v_j with different effort at random is given as

$$P_-^+(v_i) = \frac{k_-^+(v_i)}{k(v_i)}, \tag{40}$$

where $k_-^+(v_i)$ is the number of neighbors of node v_i that employ the high effort and $k(v_i)$ denotes the degree of node v_i. Since for the macroscopic description the pairwise microscopic interactions between nodes are approximated by the average dynamics, we compute the average probability P_-^+ for a node v_i with low effort to interact with a node employing the high effort. Since the network is initialized as an Erdős-Rényi random network and it is further equally likely for all nodes with the same effort to connect to or disconnect from other nodes by random rewiring, we perform a heterogeneous mean-field approximation and assume the degree $k(v_i)$ to be the same for all nodes with low effort, $k(v_i) = K_- \ \forall\, i \in \{1,\ldots,N|E_i = E_-\}$ [31,32]. Thus

$$\begin{aligned} P_-^+ &= \langle P_-^+(v_i)|E_i = E_-\rangle_i = \left\langle \frac{k_-^+(v_i)}{k(v_i)} \middle| \, E_i = E_- \right\rangle_i \\ &= \left\langle \frac{k_-^+(v_i)}{K_-} \middle| E_i = E_- \right\rangle_i = \frac{K_-^+}{K_-} \\ &= \frac{M_{+-}}{2M_- + M_{+-}}. \end{aligned} \tag{41}$$

Instead of the actual number of M links in the network we define the corresponding per node link density

$$\begin{aligned} m &= \frac{M}{N} = \frac{M_{+-}}{N} + \frac{M_-}{N} + \frac{M_+}{N} \\ &= \frac{\bar{k}}{2} = m_{+-} + m_- + m_+, \end{aligned} \tag{42}$$

which is independent of the number of nodes N. $\bar{k}$ denotes the average degree of nodes in the network, which is set to $\bar{k} = 20$ in accordance with the numerical simulations. The probability for a node with low (high) effort to interact with a node of high (low) effort is then given by

$$P_-^+ = \frac{m_{+-}}{2m_- + m_{+-}}, \tag{43}$$

$$P_+^- = \frac{m_{+-}}{2m_+ + m_{+-}}, \tag{44}$$

and is fully determined by the per node densities of links m_{+-}, m_+, and m_-.

Generally, the time evolution of the total number of links between nodes of low effort is governed by imitation and adaptation. First, we focus on the process of adaptation. Since links between nodes of the same effort can only be established but not removed via the process of adaptation, the contribution of this process to the total number of links between low-effort nodes M_- only causes it to increase. This positive contribution is

$$\frac{dM_-}{dt} \sim \rho N_- P_-^+ \tag{45}$$

and is explained as follows: In each time interval $(t, t + dt)$ there is a total number of N_- nodes which with probability ρ initiate an interaction that leads to adaptive rewiring. Adaptive rewiring then takes place if a randomly drawn neighbor v_j of the considered node v_i employs a high effort. As defined above, this happens with probability P_-^+.

The second contribution to the time evolution of M_- is given by imitation, which takes place at rate τ. Generally, there is one term causing an increase in links between nodes with low effort and one term causing its decrease. First, assume a node v_i with $E_i = E_+$ to imitate the low effort E_- from one of its neighboring nodes v_j with $E_j = E_-$. The number of links between nodes of low effort then increases by the number $k_+^-(v_i)$ of all neighbors of node v_i that employ the low effort (Fig. 4). Again, by performing a heterogeneous mean-field approximation and assuming the number of neighbors for individual nodes to be represented by the respective average

Reprinted paper with permission from M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 91, no. 5, pp. 1-11, 2015, doi: 10.1103/PhysRevE.91.052801. Copyright 2015 by the American Physical Society.

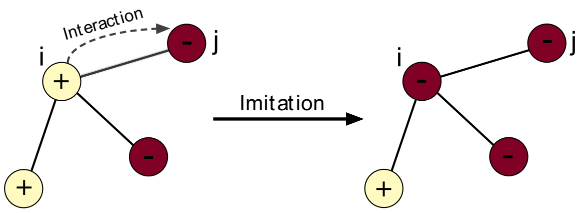


FIG. 4. (Color online) Illustration of the influence of the imitation of effort on the different numbers of link types in the network. A node v_i with the high effort $E_i = E_+$ [indicated in yellow (light)] interacts with a node v_j with low effort $E_j = E_-$ [red (dark)]. Node v_i may then imitate the effort of node v_j, $E_i \to E_-$. The number of links between nodes with low (high) effort M_- (M_+) then increases (decreases) by the number $k_+^-(v_i)$ [$k_+^+(v_i)$] of neighbors of v_i that show the low (high) effort.

number of neighbors, we set

$$k_+^-(v_j) = K_+^- = \frac{M_{+-}}{N_+}. \tag{46}$$

Now, each of the N_+ nodes with high effort interacts with a node of low effort with probability P_+^- at rate τ. Then, with probability $p_{+\to-}$, a node with high effort takes up the low effort. This causes the number of links between pairs of nodes with low effort to increase by the number of neighbors with low effort of the formerly high-effort node,

$$\frac{dM_-}{dt} \sim \tau N_+ P_+^- p_{+\to-} K_+^-. \tag{47}$$

A third term that governs the time evolution of M_- is given by its decrease caused by nodes with low effort that imitate the high effort. If a node v_i with the low effort $E_i = E_-$ interacts with a node v_j having the high effort $E_j = E_+$ and v_i then imitates the effort of v_j, the total number of links connecting two nodes with low effort decreases by the number of v_i's neighbors v_k that are showing the low effort $E_k = E_-$ as well. Following from an analogous argument as given above, this number is given by $k_-^-(v_i)$. Again we assume the number of neighbors v_k with $E_k = E_-$ of a node v_i with $E_i = E_-$ to be approximated by its average,

$$k_-^-(v_j) = K_-^- = \frac{2M_-}{N_-}. \tag{48}$$

With rate τ each of the N_- nodes with low effort interacts with a node showing the high effort E_+ with probability P_-^+. With probability $p_{-\to+}$ a node with low effort imitates the high effort which causes a decrease in M_- by the average number of low-effort neighbors K_-^- of the node that is imitating the high effort,

$$\frac{dM_-}{dt} \sim -\tau N_- P_-^+ p_{-\to+} K_-^-. \tag{49}$$

Putting together Eqs. (45), (47), and (49) gives the time evolution of the number of links between nodes of low effort as

$$\begin{aligned}\frac{dM_-}{dt} &= \tau(N_+ P_+^- p_{+\to-} K_+^- - N_- P_-^+ p_{-\to+} K_-^-)\\ &\quad + \rho N_- P_-^+.\end{aligned} \tag{50}$$

Plugging the definitions of K_-^- [Eq. (34)] and K_+^- [Eq. (38)] into Eq. (50) and normalizing with the total number of nodes N yields the time evolution of the per node density of links between nodes of low effort,

$$\frac{dm_-}{dt} = \tau(P_+^- p_{+\to-} m_{+-} - 2P_-^+ p_{-\to+} m_-) + \rho n_- P_-^+, \tag{51}$$

which is again independent of N. Due to the symmetry of the system, the time evolution of the per node density m_+ of links between nodes with high effort then immediately follows as

$$\frac{dm_+}{dt} = \tau(P_-^+ p_{-\to+} m_{+-} - 2P_+^- p_{+\to-} m_+) + \rho n_+ P_+^-. \tag{52}$$

For the time evolution of the average stock of nodes with low and high effort μ_- and μ_+ we already found in Eqs. (9) and (10) that

$$d\mu_- = dt(\mu_-(1 - \mu_- - E_-) - \mu_-^{(2)}) + \delta_-, \tag{53}$$

$$d\mu_+ = dt(\mu_+(1 - \mu_+ - E_+) - \mu_+^{(2)}) + \delta_+. \tag{54}$$

The general forms of δ_- and δ_+ are [see Eq. (12) and (13)]

$$\delta_- = \frac{\mu_+ - \mu_-}{n_-} dn_{+\to-}, \tag{55}$$

$$\delta_+ = \frac{\mu_- - \mu_+}{n_+} dn_{-\to+}. \tag{56}$$

For the case of an adaptive network, $dn_{+\to-}$ ($dn_{-\to+}$) is given by the first (second) term in Eq. (29):

$$\delta_- = \frac{\mu_+ - \mu_-}{n_-} \tau n_+ P_+^- p_{+\to-}, \tag{57}$$

$$\delta_+ = \frac{\mu_- - \mu_+}{n_+} \tau n_- P_-^+ p_{-\to+}, \tag{58}$$

with the probabilities P_-^+ and P_+^- [Eqs. (43) and (44)] as defined above and $p_{+\to-}$ and $p_{-\to+}$ being the same as for the static model [Eqs. (7) and (8)].

To summarize, the set of five coupled differential equations that represent the adaptive network model's macroscopic dynamics is given as

$$\frac{dn_-}{dt} = \tau(n_+ P_+^- p_{+\to-} - n_- P_-^+ p_{-\to+}), \tag{59}$$

$$\frac{dm_-}{dt} = \tau(P_+^- p_{+\to-} m_{+-} - 2P_-^+ p_{-\to+} m_-) + \rho n_- P_-^+, \tag{60}$$

$$\frac{dm_+}{dt} = \tau(P_-^+ p_{-\to+} m_{+-} - 2P_+^- p_{+\to-} m_+) + \rho n_+ P_+^-, \tag{61}$$

$$\frac{d\mu_-}{dt} = \mu_-(1 - \mu_- - E_-) + \tau\frac{n_+}{n_-}(\mu_+ - \mu_-)P_+^- p_{+\to-}, \tag{62}$$

$$\frac{d\mu_+}{dt} = \mu_+(1 - \mu_+ - E_+) + \tau\frac{n_-}{n_+}(\mu_- - \mu_+)P_-^+ p_{-\to+}. \tag{63}$$

Reprinted paper with permission from M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 91, no. 5, pp. 1-11, 2015, doi: 10.1103/PhysRevE.91.052801. Copyright 2015 by the American Physical Society.

It is important to note that in most previous works on adaptive networks a closed set of macroscopic equations is obtained by assuming that links in the network are drawn at random and interactions take place between nodes that are connected by them [6,33]. In this work nodes, not links, are randomly drawn and initiate an interaction with neighboring nodes. This subtle difference changes the effective time scale of the system. Specifically, in our model only a maximum of N out of all M links are affected by interactions between nodes during the same time as all M links would be considered if interactions take place by randomly drawing links in the network. In other words, in our model it takes M/N times longer to achieve the same number of updates, as one would obtain by considering per-link interactions.

For the above system, the stable fixed point's n_{-0} component can be obtained numerically for different combinations of ϕ and T [Fig. 3(b)]. The results are again in good agreement with the numerical simulations and imply that for every choice of $T > 0$ there actually exists an appropriate choice of $\phi \in [\phi_{c_1},\phi_{c_2}]$ so all nodes are likely to adopt the effort level E_-. The lower bound of the optimal rewiring probability ϕ_{c_1} can be obtained by utilizing Eq. (21) and the linear relationship between ϕ_{c_1} and T for a fixed rate of social updates τ that lead to imitation as given in Eq. (30). We thus find $\phi_{c_1}(T,\Delta E) = \phi_{c_2}(1 - \frac{2-2\Delta E^2}{1+\Delta E^2}T)\ \forall\ 0 < T < \frac{1+\Delta E^2}{2-2\Delta E^2}$ and $\phi_{c_1}(T,\Delta E) = 0$ otherwise. The upper bound ϕ_{c_2} at which the network fragments is obtained from a numerical bifurcation analysis as $\phi_{c_2} \approx 0.89$. The result is in good agreement with previous findings on the fragmentation threshold in adaptive networks for similar average degree $\bar{k}$ [34,35]. We find, however, that the computed fragmentation threshold ϕ_{c_2} is larger than what is expected from the numerical simulations [Fig. 3(a)]. This can either be due to the fact that moment closure as well as mean-field approximations are known to provide only rough estimates of the fragmentation threshold [33] or because finite-size effects in the numerical simulations cause the system to fragment for smaller values of ϕ than it would be expected for the limiting case $N \to \infty$ that is considered in the macroscopic approximations. A more detailed study of the network fragmentation and the corresponding threshold ϕ_{c_2} is a subject of future research.

C. Consistency between approximations

To illustrate the consistency of the set of differential equations describing the static setting (16)–(18) and the adaptive case (59)–(63), we set $\phi = 0$ in the latter, compute its fixed points numerically and compare them with the static setting's fixed points (21) and (22). Figures 5(a)–5(c) show the different components of the stable fixed points as a function of the control parameter T for a fixed $\Delta E = 0.5$. The components n_{-0}, μ_{+0}, and μ_{-0} align perfectly well for the static and the adaptive case. The gray shaded area in Fig. 5(a) indicates the center manifold (28) for which the system's stability cannot be assessed by standard linear stability analysis. However, numerically integrating the set of differential equations yields the expected behavior of $n_-(0) \to 0$ as $T \to 0$. Figure 5(d) displays again the n_{-0} component of the adaptive model's stable fixed point for $\phi = 0$ and different combinations of T and ΔE. The results match those of Fig. 1(b). Hence, the system of dynamic equations (59)–(63) can be interpreted as a consistent generalization of Eqs. (16)–(18).

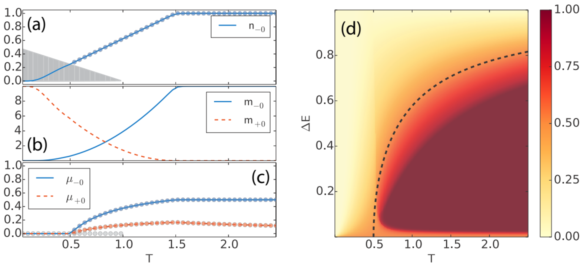


FIG. 5. (Color online) [(a)–(c)] The dependence of the adaptive (solid lines) and static model's (transparent scatter) stable fixed point on the expected waiting time T for fixed parameters $\phi = 0$ and $\Delta E = 0.5$. (d) The adaptive model's stable fixed point's n_{-0} component indicating the fraction of nodes with effort E_- in the consensus state as a function of the two parameters T and ΔE for $\phi = 0$. The dashed line indicates the value of the critical waiting time T_c obtained from the set of differential equations (16)–(18).

V. CONCLUSIONS

We have introduced a model to describe emerging structure formation from the interplay of dynamics of and on networks manifested by the coevolution of social dynamics on the one hand and resource dynamics on the other hand. An adaptive voter model has been coupled to a set of logistic growth models, such that the state of the dynamic variables influences the imitation (i.e., social trait adoption) processes in the underlying social network which take place according to differences in harvest or payoff. We have derived rate equations for the system's macroscopic variables and demonstrated that the resulting system of differential equations yields stable fixed points which are in good agreement with the results from numerical simulations.

Our paradigmatic example illustrates that the interplay between both types of network dynamics gives rise to a variety of new phenomena, which have not been observed so far when only studying either of the two aspects. We have mainly found that the rate of interactions in the network determines the expected linear stability of the growth model's fixed points. However, for each choice of interaction rate there exists an appropriate range of the adaptive rewiring frequency so that the expected fraction of, e.g., nodes with effort E_- can be maximized. Notably, the subset of differential equations (59)–(61) provides a general description of imitation and adaptation dynamics on a social network with binary states of nodes and symmetric imitation rules. Hence, it is applicable to study many other problems as long as the imitation probabilities $p_{-\to+}$ and $p_{+\to-}$, which do not have to be constant for all times, are chosen appropriately.

The proposed model also raises questions that need to be addressed in future research. In the course of the macroscopic approximation we have assumed all moments of higher order in stocks and network structure to vanish such that the set of differential equations could be closed. The results have been shown to be in good agreement with numerical simulations.

Reprinted paper with permission from M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 91, no. 5, pp. 1-11, 2015, doi: 10.1103/PhysRevE.91.052801. Copyright 2015 by the American Physical Society.

However, a more in-depth analysis of whether the inclusion of higher-order moments would enable us to reproduce the steep transition between the two regimes of predominance of low- or high-effort nodes remains a relevant research questions. We also aim to estimate more thoroughly the critical waiting time T_c at which the observed phase transition takes place and therefore investigate the expected time at which the low effort provides more harvest than the high effort given that no interaction between the nodes took place so far. Finally, we aim to obtain data from agricultural studies on, e.g., water usage or harvest exploitation of resources to test the findings and insights that we have obtained from our coevolutionary model with respect to real-world phenomena.

ACKNOWLEDGMENTS

This work was carried out within the framework of PIK's COPAN project. M.W. was supported by the German Federal Ministry for Science and Education via the BMBF Young Investigators Group CoSy-CC^2 (Grant No. 01LN1306A). J.F.D. and W.L. acknowledge funding from the Stordalen Foundation (Norway) via the PB.net initiative and BMBF (project GLUES) and J.K. acknowledges the IRTG 1740 funded by Deutsche Forschungsgesellschaft (DFG) (Germany) and FAPESP. We thank R. V. Donner for helpful comments and suggestions on the manuscript and R. Grzondziel and C. Linstead for help with the IBM iDataPlex Cluster at the Potsdam Institute for Climate Impact Research.

[1] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys. **74**, 47 (2002).

[2] M. E. J. Newman, The structure and function of complex networks, SIAM Rev. **45**, 167 (2003).

[3] T. Gross and B. Blasius, Adaptive coevolutionary networks: A review, J. R. Soc. Interface **5**, 259 (2008).

[4] T. Gross and H. Sayama, *Adaptive Networks* (Springer, Berlin, 2009).

[5] P. Holme and J. Saramäki, Temporal networks, Phys. Rep. **519**, 97 (2012).

[6] T. Gross, Carlos J. Dommar D'Lima, and B. Blasius, Epidemic dynamics on an adaptive network, Phys. Rev. Lett. **96**, 208701 (2006).

[7] R. M. May and A. L. Lloyd, Infection dynamics on scale-free networks, Phys. Rev. E **64**, 066112 (2001).

[8] R. Pastor-Satorras and A. Vespignani, Epidemic spreading in scale-free networks, Phys. Rev. Lett. **86**, 3200 (2001).

[9] G. C. M. A. Ehrhardt, M. Marsili, and F. Vega-Redondo, Phenomenological models of socioeconomic network dynamics, Phys. Rev. E **74**, 036106 (2006).

[10] P. Holme and M. E. J. Newman, Nonequilibrium phase transition in the coevolution of networks and opinions, Phys. Rev. E **74**, 056108 (2006).

[11] L. E. Blume, The statistical mechanics of strategic interaction, Game Econ. Behav. **5**, 387 (1993).

[12] G. Szabó and C. Tőke, Evolutionary prisoner's dilemma game on a square lattice, Phys. Rev. E **58**, 69 (1998).

[13] A. Traulsen, D. Semmann, R. D. Sommerfeld, H.-J. Krambeck, and M. Milinski, Human strategy updating in evolutionary games, Proc. Natl. Acad. Sci. USA **107**, 2962 (2010).

[14] P. Ji, Thomas K. DM. Peron, P. J. Menck, F. A. Rodrigues, and J. Kurths, Cluster explosive synchronization in complex networks, Phys. Rev. Lett. **110**, 218701 (2013).

[15] A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno, and C. Zhou, Synchronization in complex networks, Phys. Rep. **469**, 93 (2008).

[16] R. Perman, *Natural Resource and Environmental Economics* (Pearson Education, Harlow, 2003).

[17] H. J. Schellnhuber, Discourse: Earth System analysis—The scope of the challenge, in *Earth System Analysis* (Springer, Berlin, 1998), pp. 3–195.

[18] H. J. Schellnhuber, 'Earth system' analysis and the second Copernican revolution, Nature **402**, C19 (1999).

[19] S. J. Lade, A. Tavoni, S. A. Levin, and M. Schlüter, Regime shifts in a social-ecological system, Theor. Ecol. **6**, 359 (2013).

[20] J. M. Anderies, S. Carpenter, W. Steffen, and J. Rockström, The topology of non-linear global carbon dynamics: From tipping points to planetary boundaries, Environ. Res. Lett. **8**, 044048 (2013).

[21] O. Kellie-Smith and P. M. Cox, Emergent dynamics of the climate-economy system in the Anthropocene, Phil. T. Roy. Soc. A **369**, 868 (2011).

[22] J. Brander and M. Taylor, The simple economics of Easter Island: A Ricardo-Malthus model of renewable resource use, Am. Econ. Rev. **88**, 119 (1998).

[23] S. Motesharrei, J. Rivas, and E. Kalnay, Human and nature dynamics (HANDY): Modeling inequality and use of resources in the collapse or sustainability of societies, Ecol. Econ. **101**, 90 (2014).

[24] M. Schlüter, R. R. J. Mcallister, R. Arlinghaus, N. Bunnefeld, K. Eisenack, F. Hölker, E. Milner-Gulland, B. Müller, E. Nicholson, M. Quaas, and M. Stöven, New horizons for managing the environment: A review of coupled social-ecological systems modeling, Nat. Resour. Model. **25**, 219 (2012).

[25] J. S. Lansing and J. N. Kremer, Emergent properties of Balinese water temple networks: Coadaptation on a rugged fitness landscape, Am. Anthropol. **95**, 97 (1993).

[26] J. S. Lansing, M. P. Cox, S. S. Downey, M. A. Janssen, and J. W. Schoenfelder, A robust budding model of Balinese water temple networks, World Archaeol. **41**, 112 (2009).

[27] J. S. Lansing, *Perfect Order: Recognizing Complexity in Bali* (Princeton University Press, Princeton, NJ, 2012).

[28] A. G. Sanfey, Social decision-making: Insights from game theory and neuroscience, Science **318**, 598 (2007).

[29] H. Ebel and S. Bornholdt, Coevolutionary games on networks, Phys. Rev. E **66**, 056118 (2002).

[30] F. A. Haight, *Handbook of the Poisson Distribution* (Wiley, New York, 1967).

[31] A. Vespignani, Modelling dynamical processes in complex socio-technical systems, Nat. Phys. **8**, 32 (2012).

[32] C. Castellano and R. Pastor-Satorras, Thresholds for epidemic spreading in networks, Phys. Rev. Lett. **105**, 218701 (2010).

[33] G. Demirel, F. Vazquez, G. A. Böhme, and T. Gross, Moment-closure approximations for discrete adaptive networks, Physica D **267**, 68 (2014).

Reprinted paper with permission from M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 91, no. 5, pp. 1-11, 2015, doi: 10.1103/PhysRevE.91.052801. Copyright 2015 by the American Physical Society.

[34] H. Silk, G. Demirel, M. Homer, and T. Gross, Exploring the adaptive voter model dynamics with a mathematical triple jump, New J. Phys. **16**, 093051 (2014).

[35] G. A. Böhme and T. Gross, Analytical calculation of fragmentation transitions in adaptive networks, Phys. Rev. E **83**, 035101 (2011).

Reprinted paper with permission from M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E - Stat. Nonlinear, Soft Matter Phys., vol. 91, no. 5, pp. 1-11, 2015, doi: 10.1103/PhysRevE.91.052801. Copyright 2015 by the American Physical Society.

Eur. Phys. J. Special Topics **228**, 2357–2369 (2019)
© EDP Sciences, Springer-Verlag GmbH Germany,
part of Springer Nature, 2019
https://doi.org/10.1140/epjst/e2019-900120-4

THE EUROPEAN PHYSICAL JOURNAL SPECIAL TOPICS

Regular Article

The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources

Fabian Geier[1,2], Wolfram Barfuss[3,4], Marc Wiedermann[1,a], Jürgen Kurths[1,4,5], and Jonathan F. Donges[3,6,b]

[1] Complexity Science, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany
[2] Department of Physics, Ludwig Maximilians University, Munich, Germany
[3] Earth System Analysis, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany
[4] Department of Physics, Humboldt University, Berlin, Germany
[5] Saratov State University, Saratov, Russia
[6] Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden

Received 19 June 2019 / Received in final form 18 July 2019
Published online 28 October 2019

Abstract. Adaptive networks are a versatile approach to model phenomena such as contagion and spreading dynamics, critical transitions and structure formation that emerge from the dynamic coevolution of complex network structure and node states. Adaptive networks have been successfully applied to study and understand phenomena ranging from epidemic spreading, infrastructure, swarm dynamics and opinion formation to the sustainable use of renewable resources. Here, we study critical transitions in contagion dynamics on multilayer adaptive networks with dynamic node states and present an application to the governance of sustainable resource use. We focus on a three-layer adaptive network model, where a polycentric governance network interacts with a social network of resource users which in turn interacts with an ecological network of renewable resources. We uncover that sustainability is favored for slow interaction timescales, large homophilic network adaptation rate (as long it is below the fragmentation threshold) and high taxation rates. Interestingly, we also observe a trade-off between an eco-dictatorship (reduced model with a single governance actor that always taxes unsustainable resource use) and the polycentric governance network of multiple actors. In the latter setup, sustainability is enhanced for low but hindered for high tax rates compared to the eco-dictatorship case. These results highlight mechanisms generating emergent critical transitions in contagion dynamics on multilayer

[a] e-mail: marcwie@pik-potsdam.de

[b] e-mail: donges@pik-potsdam.de

adaptive networks and show how these can be understood and approximated analytically, relevant for understanding complex adaptive systems from various disciplines ranging from physics and epidemiology to sociology and global sustainability science. The paper also provides insights into potential critical intervention points for policy in the form of taxes in the governance of sustainable renewable resource use that can inform more process-detailed social-ecological modeling.

1 Introduction

Adaptive networks are a flexible approach to model phenomena such as contagion and spreading phenomena, critical transitions and structure formation that emerge from the dynamic coevolution of complex network structure and node states [28,36,38]. Adaptive networks have been successfully applied to study and understand phenomena ranging from epidemic spreading [37] and early warning signals for critical transitions therein [22], swarm dynamics [8,12], evolution of autocatalytic sets [31,32], opinion formation [28] and spreading of behaviors such as smoking [5] to the sustainable use of renewable resources and modelling social-ecological transformations [10,24,34,40]. Recently, adaptive dynamics have also been studied in multilayer network models that allow for representing different types of nodes or agents and their complex interconnections in a structured way [4,29,33].

Adaptive networks are also recognized as a promising approach to build a bridge between theoretical physics and efforts to understand future trajectories of the Earth system in the Anthropocene where human social dynamics have become a dominant geological process [20,27]. By modelling complex social systems as adaptive multilayer networks embedded in land-use [1] or more comprehensive Earth system models [11,19], methods from complex systems theory, nonlinear dynamics and statistical physics can be applied to identify management options, critical transitions, tipping points and critical intervention points towards sustainable development [23], map out safe operating spaces for these systems [14,16,18], and more generally, analyze complex co-evolutionary dynamics of human-environment systems including the evolution of technological and knowledge systems [6,21]. Important recent challenges in this field include the identification of sensitive intervention points for policy [13] and adaptive multi-level governance strategies [15] that can help to overcome systemic blockages and initiate the deep social-ecological transformations [34] needed to avoid dangerous anthropogenic climate change and degradation of biosphere integrity [41].

While the control of adaptive network dynamics has already been studied in the context of opinion formation influenced via zealotry [26], a stylized form of lobbyism, there is an increasing interest in studying modern polycentric, adaptive and multi-level governance and management of social–ecological systems from a complex systems perspective [15], creating bridges to the theory of governance networks from political science [7]. In this paper, we derive and analyze an adaptive multilayer network model to investigate the dynamics of an adaptive and polycentric governance network interacting with an adaptive social network of users of private renewable resources, extending upon the recently proposed and studied copan:EXPLOIT model [24,40]. In the extended model, termed copan:TAXPLOIT in this paper, governance nodes can either penalize associated unsustainable resource users by introducing taxes or can be indifferent to the level of resource exploitation (i.e., by not introducing any tax). The trait of enforcing such an environmental tax can spread contagiously on the governance network via social learning. Additionally, the governance network can adapt via homophilic rewiring. Analogously, in the resource user layer the trait of sustainable or unsustainable resource exploitation can spread via

Reprinted by permission from Springer Nature: Springer Nature, The European Physical Journal Special Topics, The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources, F. Geier et al., doi: 10.1140/epjst/e2019-900120-4, Copyright EDP Sciences, Springer-Verlag GmbH Germany, part of Springer Nature, 2019.

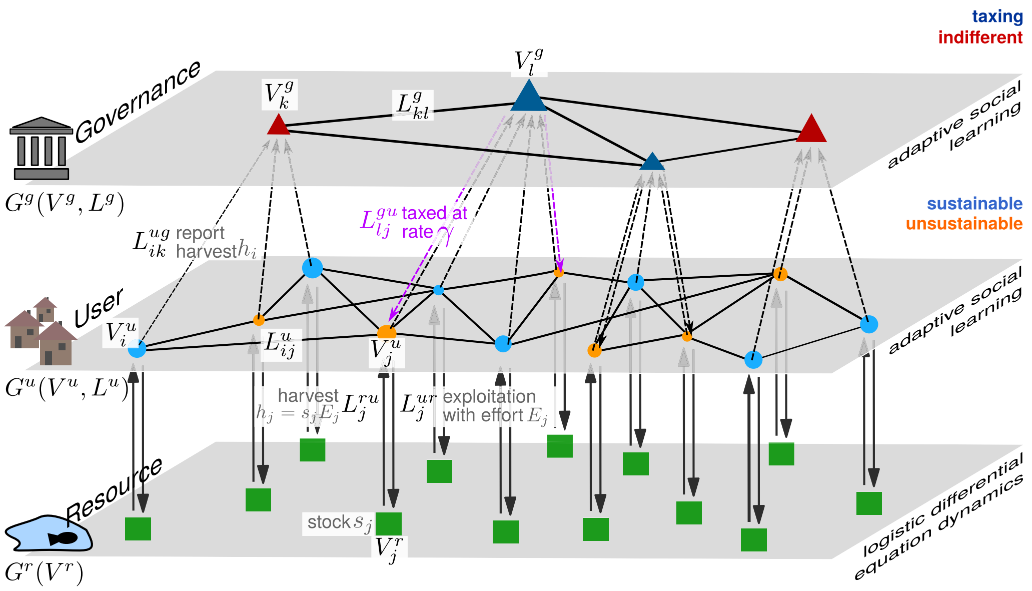


Fig. 1. Schematic visualization of the interdependent three-layer model consisting of a resource layer G^r, a user layer G^u and a governance layer G^g.

social learning and the users' social network can adapt via homophilic rewiring as well (Fig. 1).

We study this multilayer adaptive network system using numerical simulations and analytical approximations. We particularly focus on analyzing the conditions under which adaptive polycentric governance fosters the sustainable use of renewable resources and increases the resilience and size of the sustainable safe operating space of the system. We also identify critical transitions and tipping points, e.g. in the tax rate parameter, that separate domains of sustainable and unsustainable outcomes in parameter space.

We introduce the studied multilayer adaptive network model in detail (Sect. 2), report and discuss results of numerical simulations and analytical approximations (Sect. 3), and finish with concluding remarks (Sect. 4).

2 Model description

In the following we describe the multilayer adaptive network model of three layers that is studied in this work. It consists of an ecological *resource layer* representing a set of logistically growing resources, a *user layer* representing a set of agents harvesting these resources, and a *governance* layer representing agents superordinate to agents in the user layer. Nodes in the governance layer can enforce taxes on certain types of user behavior. The general setup of the model is summarized and visualized in Figure 1.

2.1 Resource layer

The first layer $G^r(V^r)$ consists of a set V^r of N_r mutually disconnected nodes $s_i,\ i = 1, \ldots, N_r$ each representing dynamics of logistic growth,

$$\frac{ds_i}{dt} = a s_i \left(1 - \frac{s_i}{K}\right), \tag{1}$$

Reprinted by permission from Springer Nature: Springer Nature, The European Physical Journal Special Topics, The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources, F. Geier et al., doi: 10.1140/epjst/e2019-900120-4, Copyright EDP Sciences, Springer-Verlag GmbH Germany, part of Springer Nature, 2019.

where a denotes the growth rate and K denotes the maximum capacity. We use identical a and K for all resource nodes for simplicity in this study, but heterogeneities in these properties can yield interesting effects as well [40]. Note that from here on, we use s_i to denote the current state of the resource but to also refer to the corresponding node in G^r. Without loss of generality we can set $a = K = 1$ and, hence, express the time t in terms of the inverse growth rate $1/a$, and the resource stock s_i in terms of the maximum capacity K. If undisturbed, s_i displays two fixed points $s_{i,0} = 0$ (unstable) and $s_{i,0} = 1$ (stable).

2.2 User layer

The second layer in our interdependent model $G^u(V^u, L^u_{ij})$ represents a set V^u of $N_u = N_r$ agents $i = 1, \ldots, N_u$ that harvest exactly one of the resources s_i with some effort (or strategy) E_i along a link $(i, s_i) \in L^{ur}_i$, where L^{ur}_i denotes the set of directed links pointing from G^u to G^r. Depending on the currently employed effort level, a node $i \in V^u$ gains an instantaneous harvest $h_i = qE_is_i$ from the resource. Here, q denotes the so-called catch coefficient or efficiency [30]. As we are only interested in an intercomparison of efforts across agents, we measure the effort in units of that efficiency by setting $q = 1$ [24]. Harvesting effectively reduces the amount of available stock s_i to each node i and hence, equation (1) is adjusted to ultimately read

$$\frac{ds_i}{dt} = s_i\,(1 - s_i) - E_is_i. \tag{2}$$

Each agent/node chooses between two values of effort level $E_- = 1 - \Delta E$ and $E_+ = 1 + \Delta E$ that cause the resource to either converge into a stable fixed point $s_{i,0} = \Delta E$ for E_- and $s_{i,0} = 0$ for E_+. We therefore denote E_- the *sustainable* and E_+ the *unsustainable* effort. In order to further reduce the number of free parameters we set $\Delta E = 0.5$ according to earlier studies [24,40] ensuring that at the fixed point $s^* = \Delta E$ the equilibrium harvest $h_0 = \Delta E(1 - \Delta E)$ is maximized.

Pairs of nodes $i \in V^u$ interact and update their strategies similarly to the adaptive voter model [4,28,29,36,38] with the process of pure imitation replaced by social learning [2,3,24,40]. Therefore, edges L^u_{ij} in the user layer G^u indicate a connection (such as friendship or business relationships) between the nodes i along which opinion formation takes place via the exchange of information on current harvesting strategies E_i and corresponding harvest h_i. To combine discrete opinion formation with continuous resource dynamics, each node is assigned a unique waiting time $T_{u,i}$ according to a Poissonian distribution that is drawn randomly after each interaction of that corresponding node i,

$$P(T_{u,i}) = \frac{1}{\Delta T_u} \exp\left(-\frac{T_{u,i}}{\Delta T_u}\right). \tag{3}$$

Here, ΔT_u is understood as the average waiting time of nodes in the user layer. It directly relates to the rate of interaction between the agents as compared to the typical timescale of the resource dynamics. In that sense, a short waiting time corresponds to more impatient agents while a high waiting time indicates comparatively patient agents.

In each time step, node i with the smallest waiting time $T_{u,i}$ becomes active and all stocks s_i are integrated forward by $T_{u,i}$. Then, a random neighbor j of i is chosen such that $(i, j) \in L^u_{ij}$. If the effort levels, i.e., strategies, E_i and E_j differ, there is a probability ϕ for i to break its connection with j and homophilically establish a new link to a formerly unconnected node n such that $E_i = E_n$. In addition, with probability $1-\phi$, i mimics the harvest strategy of j with a probability $P(E_i \to E_j)$ depending on the difference in immediate harvest h_i and h_j,

$$P(E_i \to E_j) = \frac{1}{2}\left(\tanh\left(h_j - h_i\right) + 1\right). \tag{4}$$

The hyperbolic function represents the monotonic increase in the likelihood for social learning with an increase in the expected harvest differences [2,3]. After finishing one step, a new waiting time for i is drawn according to Eq. (3) and added to the current $T_{u,i}$. This iteration scheme continues until the model reaches a consensus state where either all nodes in G^u follow the same strategy $E_i = E_j \ \forall \ i,j = 1,\dots N_u$ or G^u has fragmented into disconnected components consisting solely of nodes with the same strategy. Overall, the user and resource layers follow the same dynamics as encoded in the copan:EXPLOIT model [24,40], given that the governance layer is in an indifferent state and, hence, exerts no influence on resource users (see below).

2.3 Governance layer

Social systems often obey a hierarchical structure [35] including, e.g., super- and subordinate agents. To incorporate such effects, our model additionally consists of a third layer $G^g(V^g, L^g_{kl})$ which, for the sake of illustration, is denoted the *governance* layer. This layer consists of N_g nodes k that are connected via a set of links L^g_{kl} indicating an abstract form of, e.g. diplomatic relationships. Nodes k can be in one of either two states S_k: S_- (*taxing*) or S_+ (*indifferent*), which are to some extent analogous to the sustainable and unsustainable states of nodes in the user layer G^u. Additionally each node $i \in V^u$ in the user layer is connected to exactly one node $k \in V^g$ in the governance layer (implying that $N_g \leq N_u$).

Nodes $k \in V^g$ also follow an opinion formation process along the lines of the extended adaptive voter model as described above. Hence, for each node $k \in V^g$ we draw waiting times $T_{g,i}$ according to equation (3) and set an average waiting time ΔT_g unique to the governance layer G^g. As above, once node k becomes active, a neighbor l that is connected with k is drawn uniformly at random. With probability ϕ and if the states of the two nodes differ (i.e., $S_l \neq S_k$), k breaks its connection with l and establishes a new link to a previously unconnected node $n \in V^g$, such that $S_k = S_n$. For the sake of reducing the number of free parameters, we employ the same rewiring probability ϕ in G^u and G^g. In contrast to nodes $i \in V^u$, a node $k \in V^g$ does not harvest from its own resource stock, but instead accumulates the harvests h_i of all nodes i that k is connected to via interdependence links L^{ug}_{ik}, such that

$$h_k = \sum_{i \in V^u | (i,k) \in L^{ug}_{ik}} h_i, \ k \in V^g. \tag{5}$$

Hence, the probability for a node $k \in V^g$ to update its state S_k to state S_l of one of its neighboring nodes l then reads,

$$P(S_k \to S_l) = \frac{1}{2}\left(\tanh\left(h_l - h_k\right) + 1\right). \tag{6}$$

As in equation (4) the hyperbolic tangent represents the experimentally observed increased likelihood for social learning as a function of the difference in cumulative harvest [2,3]. If k is now in the taxing state, it favors the long-term sustainable strategy E_- and, hence, taxes those connected subordinate nodes $i \in V^u$ that are employing the non-sustainable strategy E_+ at a rate $\gamma \in [0,1]$. γ effectively lowers the harvests of nodes i with $E_i = E_+$ such that the probability for learning another node's effort level as given in equation (4) is modified to read

$$P\left(E_i \rightarrow E_j\right) = \frac{1}{2}\tanh\left(\alpha_j h_j - \alpha_i h_i\right) + \frac{1}{2} \tag{7}$$

where $\alpha_i = (1-\gamma)$ if $E_i = E_+$ and the superordinate node $k \in V^g$ of i is in the taxing state (the same holds for node j). Otherwise, we set $\alpha_i = 1$ ($\alpha_j = 1$). Thus, governance nodes k in the taxing state punish unsustainable strategies of nodes i in the user layer.

In order to ensure that the social learning process in both layers reaches consensus at approximately the same time T_c we demand that

$$T_c = X_g \Delta T_g = X_u \Delta T_u, \tag{8}$$

where X_g and X_u is the total number of pairwise interactions between nodes in G^g and G^u, respectively. Assuming that only learning, no adaptation and no interaction between the layers takes place, $X_\bullet$ has previously been analytically derived as $X_\bullet = N\mu_1^2/\mu_2$ [39], where μ_n is the nth moment of the degree distribution and N is the number of nodes in the respective network. As we initialize each layer as an Erdős–Renyí random graph [25] with linking probability ρ (see Sect. 2.4) we obtain a Poissonian degree distribution with $\mu_1 = \mu_2 = N\rho$. This yields

$$N_g^2 \rho \Delta T_g = N_u^2 \rho \Delta T_u \rightarrow \Delta T_g = (N_u/N_g)^2 \Delta T_u. \tag{9}$$

Hence, we can express the average waiting time ΔT_g for nodes $k \in V^g$ in the governance layer in terms of the average waiting time ΔT_u for nodes $i \in V^u$ in the user layer G^u. This assumption also holds if one considers an adaptive network where only rewiring and no change in node state ($\phi = 1$) takes place. Then the time to reach a fragmented state depends linearly on the number of edges that connect nodes of different states. If one considers a random network topology with only two uniformly distributed node states, this number of cross-links (and thus $X_\bullet$) again depends quadratically on the number of nodes such that equation (9) also holds for this limiting case. A further in-depth investigation of $X_\bullet$ for cases of $\phi \in (0,1)$ is beyond the scope of this work. However, for the purpose of dimension reduction within this study we assume equation (9) to approximately hold for those cases as well.

Also note, since we demanded $N_g \leq N_u$ it follows that $\Delta T_g \geq \Delta T_u$, which is consistent with the association of network layers to users and governance actors such that governance processes commonly happen on a slower timescale than economic resource use decisions.

2.4 Initial conditions and model setup

For the following analysis we initialize our model as three coupled Erdős–Renyí random graphs with $N_u = N_r = 500$, $N_g = 50$, linking probability $\rho_g = \rho_u = 0.05$ for the governance (G^g) and the user layer (G^u), and linking probability $\rho_r = 0$ for the resource layer (G^r). Each node $i \in V^u$ in the user layer is connected to exactly one randomly drawn node $k \in V^g$ in the governance layer. All stocks in the resource layer are initially set to $s_i(t=0) = 1,\ \forall\ i = 1, \ldots, N_r$. For each node in the user layer, an initial effort of E_+ or E_- is drawn uniformly at random. The same holds (if not specified otherwise) for the initial states of nodes in the governance layer. For each combination of the taxrate γ, rewiring probability ϕ, and average waiting time in the user layer ΔT_u, we perform Monte-Carlo simulations with $M = 100$ ensemble members until at least the user layer reaches its consensus state.

Reprinted by permission from Springer Nature: Springer Nature, The European Physical Journal Special Topics, The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources, F. Geier et al., doi: 10.1140/epjst/e2019-900120-4, Copyright EDP Sciences, Springer-Verlag GmbH Germany, part of Springer Nature, 2019.

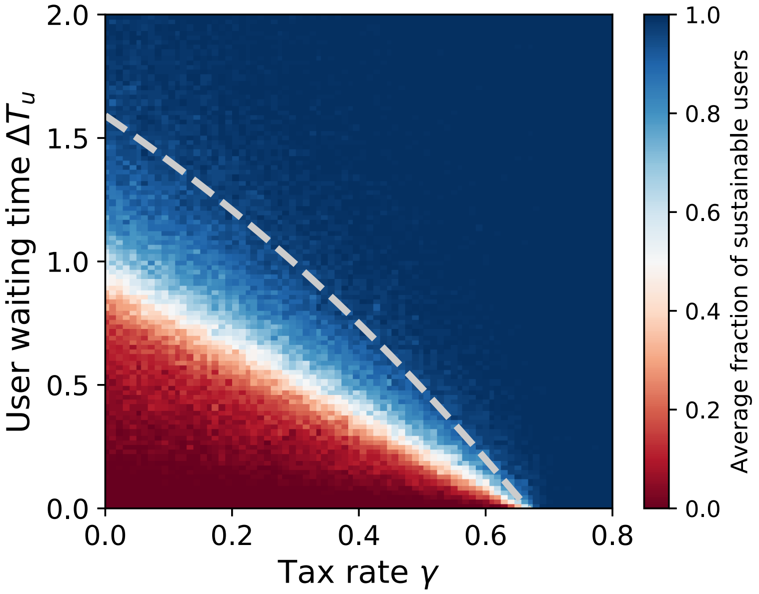


Fig. 2. Average fraction of sustainable users when unsustainable nodes are always taxed at rate γ and no rewiring takes place ($\phi = 0$). The dashed line gives the macroscopic critical time $\Delta T_{u,\mathrm{crit}}$ as given in equation (12). $\Delta T_{u,\mathrm{crit}}$ decreases with increasing γ such that for a sufficiently large taxation the critical update time approaches zero at γ_{crit}.

3 Results and discussion

3.1 Social learning in the user layer

We start the analysis by considering a governance network G^g with only one node that is in the taxing state. This means that effectively no learning dynamics take place in the governance layer and all nodes in the user layer that employ the unsustainable strategy (effort level E_+) are automatically taxed at rate γ. We refer to this setup as an "eco-dictatorship" in the following. Additionally, we first focus on the case with no adaptation in either layer and hence, set $\phi = 0$. Thus, we focus on a case with solely social learning, i.e., an imitation of harvesting strategies, in the user layer. The corresponding average fraction of sustainable users in the consensus state depending on the choice of tax rate γ and user waiting time ΔT_u is displayed in Figure 2. We mainly find that, for low tax rates γ and low user waiting times ΔT_u the system is most likely to converge into a consensus state with all nodes employing the unsustainable strategy (lower left corner of Fig. 2). This is caused by the fact that for low values of ΔT_u most pairwise interactions take place before the resource stocks of the unsustainable agents are depleted to a state where they yield less harvest than those stocks of sustainable agents. In other words, for low ΔT_u the interaction time-scale becomes much shorter than the time-scale of resource dynamics. With increasing tax rate γ the system converges more likely into a sustainable state even at comparatively low user waiting times ΔT_u as the effective harvest of unsustainable agents is reduced more drastically. For very large tax rates γ and/or very large user waiting times ΔT_u the system converges into a sustainable state as the resource stocks of unsustainable agents are close to their stable fixed point at $s^* = 0$ when most pairwise interactions happen. On the other hand, high tax rates γ further decrease the effective harvest of unsustainable agents such that an imitation of the unsustainable strategy becomes less and less likely (Fig. 2). We additionally observe that there exists a critical user waiting time $\Delta T_{u,\mathrm{crit}}$ above which the system always converges into a sustainable state regardless of the choice of tax rate γ. The same

Reprinted by permission from Springer Nature: Springer Nature, The European Physical Journal Special Topics, The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources, F. Geier et al., doi: 10.1140/epjst/e2019-900120-4, Copyright EDP Sciences, Springer-Verlag GmbH Germany, part of Springer Nature, 2019.

holds for γ itself as there seems to exist a critical value γ_{crit} above which the system also very likely converges into a sustainable state. In the following we aim to estimate values of $\Delta T_{u,\mathrm{crit}}$ and γ_{crit}.

3.2 Analytical treatment of limiting cases

For the eco-dictatorship setup studied above, we approximate a critical update time $\Delta T_{u,\mathrm{crit}}$ at which the sustainable strategy E_- becomes profitable in terms of immediate harvest when compared to the unsustainable one (E_+). For this we assume that agents do not update their strategy at times $t < \Delta T_{u,\mathrm{crit}}$ and hence $E_i(t) = E_i(0)\ \forall\ t < \Delta T_{u,\mathrm{crit}}$. In this case the temporal evolution of the corresponding stocks is obtained by integrating equation (2) which yields:

$$s^{\pm}(t) = \frac{\mp\Delta E}{(\mp\Delta E - 1)e^{\pm\Delta E t} + 1}. \tag{10}$$

Here $s^+(t)$ ($s^-(t)$) denotes the stock for those agents that employ the unsustainable (sustainable) strategy. If no interactions take place the two strategies yield the same harvest $h_\bullet(\Delta T_{u,\mathrm{crit}})$ at time $t = \Delta T_{u,\mathrm{crit}}$ and, hence,

$$(1-\gamma)s^+(\Delta T_{u,\mathrm{crit}})(1+\Delta E) = s^-(\Delta T_{u,\mathrm{crit}})(1-\Delta E). \tag{11}$$

Plugging in equation (10) yields

$$(1-\gamma)e^{-\Delta E\Delta T_{u,\mathrm{crit}}} + e^{\Delta E\Delta T_{u,\mathrm{crit}}} = \frac{2-\gamma-\gamma\Delta E}{1-\Delta E^2}. \tag{12}$$

This equation of the general form $ae^{-x} + e^x = b$ is solved by using $x = \ln\left(\frac{1}{2}\left(b \pm \sqrt{b^2-4a}\right)\right)$. Figure 2 shows the critical user waiting time $\Delta T_{u,\mathrm{crit}}$ as a function of the tax rate γ. We find that the approximated functional form of $\Delta T_{u,\mathrm{crit}}(\gamma)$ provides a boundary above which the sustainable strategy almost always succeeds. As expected $\Delta T_{u,\mathrm{crit}}$ approaches zero with increasing γ as higher tax rates reduce the effective harvest of unsustainable agents and, hence, makes the sustainable harvest profitable much earlier in time.

The critical tax rate γ_{crit}, beyond which the sustainable resource use is always maintained (for $\gamma \geq \gamma_{\mathrm{crit}}$), can be derived by setting $\Delta T_{u,\mathrm{crit}} = 0$ in equation (12). This yields

$$\gamma_{\mathrm{crit}}(\Delta E) = 2\frac{\Delta E}{1+\Delta E} \tag{13}$$

and $\gamma_{\mathrm{crit}} = \frac{2}{3}$ for $\Delta E = 0.5$. This matches the numerically computed result of $\gamma_{\mathrm{crit}} \approx 0.67$ (Fig. 2, intersection of dashed grey line with the abscissa $\Delta T_u = 0$).

3.3 Social learning in governance and user layer

After obtaining the results for a simplified governance layer with just one single node (the eco-dictatorship), we now turn to the analysis of the full model by setting $N_g = 50$ and, hence, allowing for social learning as described in Section 2.3 in the governance layer as well. Additionally we allow for adaptive rewiring by setting the rewiring probability to an intermediate value of $\phi = 0.4$ (Fig. 3). Recall that this implies that whenever two nodes of different state or strategy in either layer interact there is a probability of ϕ for the link between those two nodes to be homophilically rewired such that two nodes of the same strategy or state are connected afterwards.

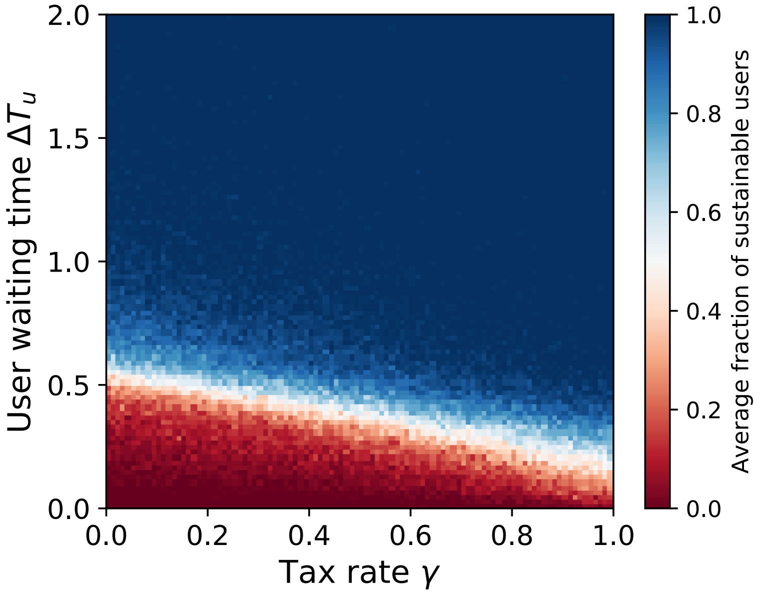


Fig. 3. If social learning happens in both, the user and the governance layer, rewiring at an intermediate probability (here with $\phi = 0.4$) induces a trade-off. In particular, the system is more likely to become sustainable at low tax rates but less likely to become sustainable at high tax rates as compared to a setting where unsustainable nodes are always taxed (compare Fig. 2).

As a first general observation, we find that increasing the tax rate γ increases the size of the sustainable regime, i.e., the system converges to a state with all nodes employing the sustainable strategy at smaller user waiting times T_u (Fig. 3). Hence, increasing the tax rate also increases the resilience of the entire system. We further observe the absence of a critical tax rate γ_{crit} and, hence, there is no γ for which the system always converges into the sustainable state (compare Figs. 2 and 3). Hence, social learning and network adaptation induce a trade-off (as compared to the case of a single sustainable governance node) where the sustainable regime increases in size for the case of small tax rates γ but decreases in size for larger tax rates γ (compare again the size of the sustainable regimes between Figs. 2 and 3). This phenomenon is explained in the following.

Since the governance layer now partly consists of nodes that are in the *indifferent* state, there is a chance for unsustainable nodes in the user layer for not being taxed. In that case, their harvest likely exceeds that of sustainable nodes in the beginning of the simulation if the average user waiting time ΔT_u is small. At the same time, even if unsustainable nodes are being taxed at a moderate rate their harvest might exceed that of sustainable nodes if their corresponding stocks are still far away from equilibrium, i.e., depletion. Hence, the increased size of the unsustainable regime at larger tax rates can be attributed to the effect of social learning in the governance layer.

For low tax rates we observe a decrease in the size of the unsustainable regime as compared to the case of no social learning in the governance layer (Sect. 3.1) and, more importantly, no network adaptation in either layer. It has been observed already in earlier studies that adaptation fosters the tendency of the system to reach the sustainable state as it allows nodes of the same strategy to form clusters [24]. This clustering of specifically the sustainable nodes allows them to avoid exposure to the unsustainable strategy (E_+) until the sustainable strategy (E_-) has become more profitable. From there on the sustainable strategy can spread through the network and tip the entire system into a sustainable state even at lower user waiting times as

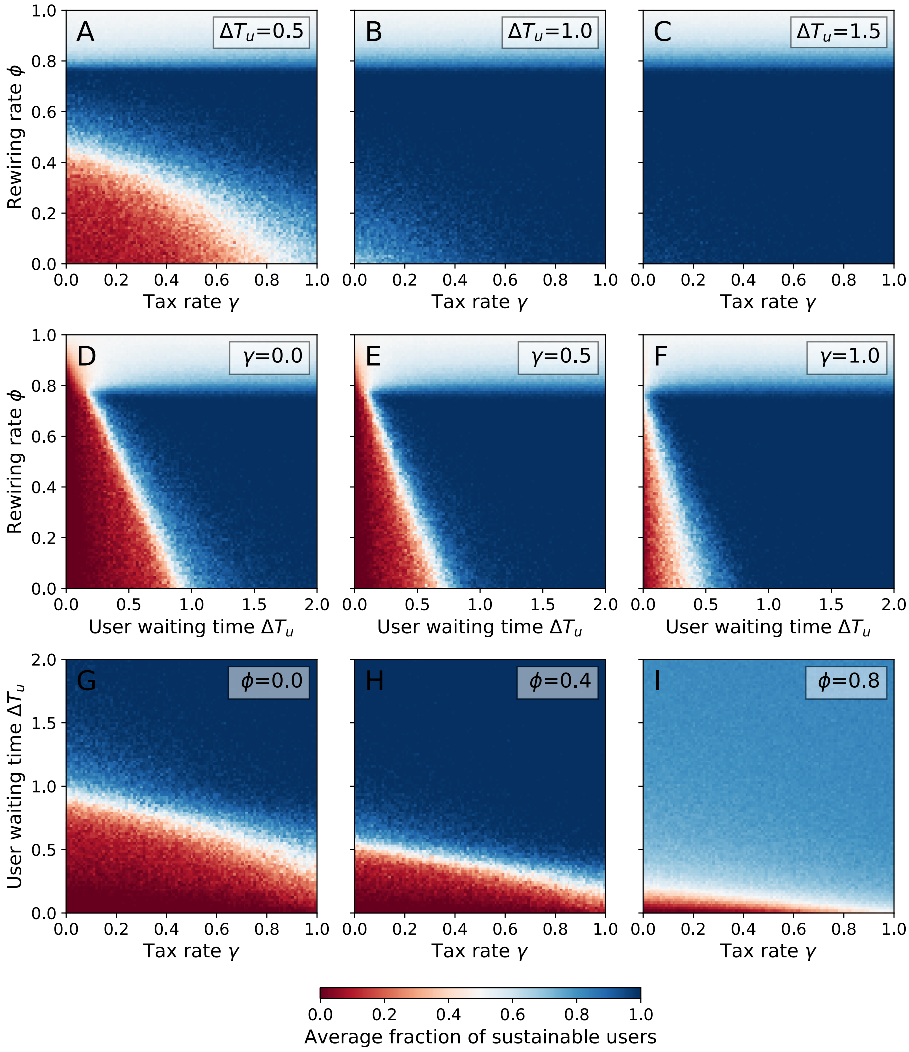


Fig. 4. The average fraction of sustainable agents for different combinations of tax rate γ, user waiting time ΔT_u and adaptation probability ϕ.

compared to the case with no adaptation (compare lower left parts of Figs. 2 and 3). Hence, the decrease in the size of the unsustainable regime for lower tax rates is mainly attributed to the presence of adaptation in both, the user and the governance layer.

In summary, we find that social learning in the governance layer increases the size of the unsustainable regime at high tax rates when compared to the case of an absence of social learning. At the same time, adaptive rewiring increases the size of the sustainable regime at lower tax rates. In other words, at low tax rates network adaptation and governmental social learning are preferred to drive the system into a sustainable state, while at high tax rates social learning and adaptation are to be avoided.

3.4 Comprehensive analysis

We ultimately vary the three crucial parameters γ (tax rate), ϕ (adaptation/rewiring probability) and ΔT_u (average user waiting time) to provide a comprehensive analysis and to illustrate how the size of the sustainable regime depends on their particular

Reprinted by permission from Springer Nature: Springer Nature, The European Physical Journal Special Topics, The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources, F. Geier et al., doi: 10.1140/epjst/e2019-900120-4, Copyright EDP Sciences, Springer-Verlag GmbH Germany, part of Springer Nature, 2019.

choices (Fig. 4). We summarize our three main results below:

(i) First, we present results for three choices of average user waiting time $\Delta T_u = 0.5$ (Fig. 4A), $\Delta T_u = 1.0$ (Fig. 4B) and $\Delta T_u = 1.5$ (Fig. 4C) and varying values of γ and ϕ. For all three cases, we first observe that there exists a fragmentation threshold at around $\phi \approx 0.8$ above which the final share of sustainable nodes in the user layer roughly corresponds to the expected initial share of 0.5 (Figs. 4A–4C). In addition we find that for $\Delta T_u = 0.5$ there exists an unsustainable regime for low tax rates γ and low rewiring probabilities ϕ as the myopic agents do not foresee a potential collapse of their respective resource stocks when being unsustainable or indifferent. However, the size of this regime decreases in size with increasing γ (Fig. 4A) as the unsustainable strategy becomes less profitable. With increasing the user waiting time to $\Delta T_u = 1.0$ (Fig. 4B) or $\Delta T_u = 1.5$ (Fig. 4C) the system converges into a sustainable state for almost all choices of γ and ϕ as long as the rewiring rate is chosen such that the fragmentation threshold is not transgressed. Hence, we conclude that the larger the average user waiting time, the more likely the system converges into a sustainable state (as it is also reported in earlier studies [24,40]).

(ii) Next, we present the results for three choices of tax rate $\gamma = 0$ (no taxation), $\gamma = 0.5$ (intermediate taxation) and $\gamma = 1$ (full taxation) and varying user waiting time ΔT_u as well as rewiring probability ϕ (Figs. 4D–4F). For the case of no taxation, i.e., no effect of the governance layer (Fig. 4D), the size of the sustainable regime increases linearly with increasing ϕ until, again, the fragmentation transition is reached. This result is in accordance with earlier studies that investigate the effect of social learning and adaptation in a system that is only comprised of the user and the resource layer [24]. Increasing the tax rate steadily decreases the size of the unsustainable regime (Figs. 4E, 4F), while the linear dependence between the rewiring probability ϕ and the size of the regime sizes persists. Remarkably, even for the case of full taxation (Fig. 4F) the unsustainable regime remains to exist as for very low user waiting times ΔT_u the unsustainable and indifferent strategies can spread through both layers as the resource stocks deplete slower compared to the rate of social interactions.

(iii) Ultimately, we consider three cases of rewiring probabilities, i.e., $\phi = 0$ (no rewiring and only social learning), $\phi = 0.4$ (intermediate rewiring) and $\phi = 0.8$ (almost only rewiring at a rate close to the fragmentation threshold and few cases of social learning), Figures 4G–4I. Note that Figure 4H shows the same results as the previously discussed Figure 3. As already discussed in Section 3.3 social learning in the governance layer causes the absence of a critical tax rate γ_{crit}, that we observed from the case of a single sustainable governance node, Figure 4G. However, allowing for rewiring at an intermediate rate (Fig. 4H) again yields an increase in the size of the sustainable regime. Further increasing the rewiring probability causes the size of the sustainable regime to increase even further. However, as the system approaches the fragmentation transition, the average fraction of sustainable users is lowered due to the formation of isolated clusters of user and governance nodes that solely employ the unsustainable/indifferent strategy (Fig. 4I).

In summary, we observe that the system is most likely to reach a sustainable regime if a high tax rate γ and a rewiring probability ϕ close to (but still below) the fragmentation transition are chosen. In other words, such a combination of parameters maximizes the size of the sustainable regime. Given that an implementation of arbitrarily high tax rates is often not feasible, minimal/optimal tax rates could be chosen for a given user waiting time ΔT_u and rewiring probability ϕ such that the system is likely to converge into a sustainable state while putting the least amount of pressure as possible onto users that show an undesired strategy (see e.g., Fig. 4A).

Reprinted by permission from Springer Nature: Springer Nature, The European Physical Journal Special Topics, The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources, F. Geier et al., doi: 10.1140/epjst/e2019-900120-4, Copyright EDP Sciences, Springer-Verlag GmbH Germany, part of Springer Nature, 2019.

4 Conclusion

In this article, we have developed a stylized model for polycentric hierarchical governance structures with a focus on investigating the preconditions for the sustainable use of renewable resources. While resource users can employ either a sustainable or non-sustainable harvesting strategy, policies are implemented via either taxation or no taxation of non-sustainable resource use. The model design is targeted towards a better systems understanding where governance actors' and resource users' interactions are driven by the following two social processes: social learning of favorable strategies and homophilic network adaptation, but take place on different hierarchical scales.

Generally we find that sustainability is favored for slow interaction timescales, large homophilic network adaptation (as long as it is below the fragmentation threshold) and high taxation rates. For the case of an eco-dictatorship, where a single governance actor taxes all non-sustainable behavior, we find the intuitive result that a sufficiently large taxation rate always causes a sustainable outcome. In contrast, in the fully process-driven model with social learning and homophilic network adaptation among governance actors, we find a trade-off: sustainability is enhanced for low and hindered for high tax rates compared to the results obtained for the eco-dictatorship.

This rather non-intuitive result highlights that the emergent outcomes of freely co-evolving social processes can be preferable compared to those obtained with a benevolent centralistic actor if low tax rates are a normative preference. In this regard, our model serves as a stylized example to find minimal tax rates that still guarantee an optimally sustainable outcome under polycentric governance structures, given a social learning process with a certain network adaptation rate and interaction timescale.

Possibly, our model could serve as a prototype for more detailed studies to be targeted at the question of optimal carbon taxes rates [9]. It highlights how social processes such as opinion formation may be combined with macro-economic optimization techniques [19] in order to gain momentum on the road to the much needed rapid decarbonization [17].

This work was conducted in the framework of the COPAN collaboration (www.pik-potsdam.de/copan) at the Potsdam Institute for Climate Impact Research (PIK). We are grateful for financial support by the Heinrich Böll Foundation, Leibniz Association (project DominoES), Stordalen Foundation (via the Planetary Boundary Research Network PB.net), the Earth League's EarthDoc program, and the European Research Council (ERC advanced grant project ERA). Parameter studies were performed on the high-performance compute cluster of PIK, supported by the European Regional Development Fund, BMBF, and the Land Brandenburg. The authors thank Jobst Heitzig and Wolfgang Lucht for inspiring discussions.

References

1. A. Arneth, C. Brown, M.D.A. Rounsevell, Nat. Clim. Change **4**, 550 (2014)
2. A. Traulsen, D. Semmann, R.D. Sommerfeld, H.-J. Krambeck, M. Milinski, Proc. Natl. Acad. Sci. USA **107**, 2962 (2010)
3. A. Traulsen, J.M. Pacheco, M.A. Nowak, J. Theor. Biol. **246**, 522 (2007)
4. B. Min, M.S. Miguel, New J. Phys. **21**, 035004 (2019)
5. C.-F. Schleussner, J.F. Donges, D.A. Engemann, A. Levermann, Sci. Rep. **6**, 30790 (2016)
6. C. Herrmann-Pillath, Ecol. Econ. **149**, 212 (2018)
7. C.J. Koliba, J.W. Meek, A. Zia, R.W. Mills, *Governance networks in public administration and public policy* (Routledge, 2018)
8. C. Huepe, G. Zschaler, A.-L. Do, T. Gross, New J. Phys. **13**, 073022 (2011)

9. D. Klenert, L. Mattauch, E. Combet, O. Edenhofer, C. Hepburn, R. Rafaty, S. Nicholas, Nat. Clim. Change **8**, 669 (2018)
10. F. Müller-Hansen, J. Heitzig, J.F. Donges, M.F. Cardoso, E.L. Dalla-Nora, P. Andrade, J. Kurths, K. Thonicke, Ecol. Econ. **159**, 198 (2019)
11. F. Müller-Hansen, M. Schlüter, M. Mäs, J.F. Donges, J.J. Kolb, K. Thonicke, J. Heitzig, Earth Sys. Dyn. **8**, 977 (2017)
12. I.D. Couzin, C.C. Ioannou, G. Demirel, T. Gross, C.J. Torney, A. Hartnett, L. Conradt, S.A. Levin, N.E. Leonard, Science **334**, 1578 (2011)
13. J.D. Farmer, C. Hepburn, M.C. Ives, T. Hale, T. Wetzer, P. Mealy, R. Rafaty, S. Srivastav, R. Way, Science **364**, 132 (2019)
14. J.-D. Mathias, J.M. Anderies, M. Janssen, Earth's Future **6**, 1555 (2018)
15. J.-D. Mathias, S. Lade, V. Galaz, Int. J. Commons **11**, 1 (2017)
16. J. Heitzig, T. Kittel, J.F. Donges, N. Molkenthin, Earth Syst. Dyn. **7**, 1 (2016)
17. J. Rockström, O. Gaffney, J. Rogelj, M. Meinshausen, N. Nakicenovic, H.J. Schellnhuber, Science **355**, 1269 (2017)
18. J.M. Anderies, J.-D. Mathias, M.A. Janssen, Proc. Natl. Acad. Sci. **116**, 5277 (2019)
19. J.F. Donges, J. Heitzig, W. Barfuss, J.A. Kassel, T. Kittel, J.J. Kolb, T. Kolster, F. Müller-Hansen, I.M. Otto, M. Wiedermann, K.B. Zimmerer, W. Lucht, Earth Syst. Dyn. Discuss. (2018), https://doi.org/10.5194/esd-2017-126
20. J.F. Donges, R. Winkelmann, W. Lucht, S.E. Cornell, J.G. Dyke, J. Rockström, J. Heitzig, H.J. Schellnhuber, Anthropocene Rev. **4**, 151 (2017)
21. J. Renn, M. Laubichler, Extended Evolution and the History of Knowledge, in *Integrated History and Philosophy of Science* (Springer, 2017), pp.109–125
22. L. Horstmeyer, C. Kuehn, S. Thurner, Phys. Rev. E **98**, 042313 (2018)
23. M. Milkoreit, J. Hodbod, J. Baggio, K. Benessaiah, R. Calderón-Contreras, J.F. Donges, J.D. Mathias, J.C. Rocha, M. Schoon, S.E. Werners, Environ. Res. Lett. **13**, 033005 (2018)
24. M. Wiedermann, J.F. Donges, J. Heitzig, W. Lucht, J. Kurths, Phys. Rev. E **91**, 052801 (2015)
25. P. Erdős, A. Rényi, *On the Evolution of Random Graphs* (Publication of the Mathematical Institute of the Hungarian Academy of Sciences, 1960), pp. 17–61
26. P.P. Klamser, M. Wiedermann, J.F. Donges, R.V. Donner, Phys. Rev. E **96**, 052315 (2017)
27. P.H. Verburg, J.A. Dearing, J.G. Dyke, S. Van Der Leeuw, S. Seitzinger, W. Steffen, J. Syvitski, Global Environ. Change **39**, 328 (2016)
28. P. Holme, M.E.J. Newman, Phys. Rev. E **74**, 056108 (2006)
29. R. Amato, N.E. Kouvaris, M.S. Miguel, A. Díaz-Guilera, New J. Phys. **19**, 123019 (2017)
30. R. Perman, *Natural resource and environmental economics* (Pearson Education, 2003)
31. S. Jain, S. Krishna, Phys. Rev. Lett. **81**, 5684 (1998)
32. S. Jain, S. Krishna, Proc. Natl. Acad. Sci. USA **98**, 543 (2001)
33. S. Boccaletti, G. Bianconi, R. Criado, C.I. Del Genio, J. Gómez-Gardenes, M. Romance, I. Sendina-Nadal, Z. Wang, M. Zanin, Phys. Rep. **544**, 1 (2014)
34. S.J. Lade, Ö. Bodin, J.F. Donges, E.E. Kautsky, D. Galafassi, P. Olsson, M. Schlüter, arXiv:1704.06135 (2017)
35. T. Parsons, *An Outline of the Social System* (University of Puerto Rico, Department of Social Sciences, 1961)
36. T. Gross, B. Blasius, J.R. Soc. Interface **5**, 259 (2008)
37. T. Gross, C.J. Dommar D'Lima, B. Blasius, Phys. Rev. Lett. **96**, 208701 (2006)
38. T. Gross, H. Sayama, *Adaptive networks* (Springer, 2009)
39. V. Sood, S. Redner, Phys. Rev. Lett. **94**, 178701 (2005)
40. W. Barfuss, J.F. Donges, M. Wiedermann, W. Lucht, Earth System Dyn. **8**, 255 (2017)
41. W. Steffen, J. Rockström, K. Richardson, T.M. Lenton, C. Folke, D. Liverman, C.P. Summerhayes, A.D. Barnosky, S.E. Cornell, M. Crucifix, J.F. Donges, I. Fetzer, S.J. Lade, M. Scheffer, R. Winkelmann, H.J. Schellnhuber, Proc. Natl. Acad. Sci. **115**, 8252 (2018)

3.4 *Model simplification and approximation methods*

In this last section of the third part of the reader we showcase developments of simplification and approximation methods to get a deeper understanding of dynamics on complex networks.

In the first paper, "Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model" [Kolb et al., 2020], we applied approximation methods known from statistical physics to derive a set of ordinary differential equations that approximate the macro-dynamics occuring in a multi-agent network model. As an exemplary model we focussed on a predominantly socio-economic one this time. Similar approximation techniques have been used by [Wiedermann et al., 2020] to derive macroscopic equations describing the spread of collective action from underlying microscopic network processes.

In a more machine-learning oriented paper, we investigated the behaviour of learning agents interacting with dynamic environments by approximating it by deterministic equations in "Deterministic limit of temporal difference reinforcement learning for stochastic games" [Barfuss et al., 2019]. We demonstrated the potential of our method with the three well-established reinforcement learning algorithms of Q-learning, SARSA learning, and actor-critic learning.

Finally, "Dynamics of tipping cascades on complex networks" [Krönke et al., 2020] focuses on the preconditions for the emergence of tipping cascades on complex networks. In particular, we studied the effects of network topology on the occurrence of such cascades.

PHYSICAL REVIEW E **102**, 042311 (2020)

Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model

Jakob J. Kolb[*]
FutureLab on Game Theory and Networks of Interacting Agents, Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany and Department of Physics, Humboldt University Berlin, 10117 Berlin, Germany

Finn Müller-Hansen
Mercator Research Institute on Global Commons and Climate Change, 10829 Berlin, Germany and Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany,

Jürgen Kurths
Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany and Department of Physics, Humboldt University Berlin, 10117 Berlin, Germany

Jobst Heitzig
FutureLab on Game Theory and Networks of Interacting Agents, Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany

(Received 4 October 2019; accepted 13 July 2020; published 28 October 2020)

In this paper, we propose a statistical aggregation method for agent-based models with heterogeneous agents that interact both locally on a complex adaptive network and globally on a market. The method combines three approaches from statistical physics: (a) moment closure, (b) pair approximation of adaptive network processes, and (c) thermodynamic limit of the resulting stochastic process. As an example of use, we develop a stochastic agent-based model with heterogeneous households that invest in either a fossil-fuel- or renewables-based sector while allocating labor on a competitive market. Using the adaptive voter model, the model describes agents as social learners that interact on a dynamic network. We apply the approximation methods to derive a set of ordinary differential equations that approximate the macrodynamics of the model. A comparison of the reduced analytical model with numerical simulations shows that the approximation fits well for a wide range of parameters. The method makes it possible to use analytical tools to better understand the dynamical properties of models with heterogeneous agents on adaptive networks. We showcase this with a bifurcation analysis that identifies parameter ranges with multistabilities. The method can thus help to explain emergent phenomena from network interactions and make them mathematically traceable.

DOI: 10.1103/PhysRevE.102.042311

I. INTRODUCTION

Agent-based modeling is a computational approach to simulate systems composed of a large number of similar subunits with many applications in ecology [1], business [2], sociology [3], and economics [4,5]. Agent-based models (ABMs) are used to study aggregate phenomena emerging from local interactions [6]. These interactions can be structured by spatial embedding of agents or by social networks [7–10]. In economics, ABMs have been used to study, for example, business cycles [11], market power [4], and trade [5].

ABMs are a promising alternative to dynamic stochastic general equilibrium (DSGE) modeling, the current workhorse of theoretical macroeconomics. DSGE models usually build on the representative agent approach, i.e., they represent all individuals of one type such as firms or consumers by one representative decision maker.

The representative agent approach implies that theoretical macroeconomics reduces macroeconomic phenomena to assumptions about a few different representative agents, leaving out many explanatory mechanisms for fluctuations in aggregate variables based on intragroup interaction and heterogeneity [12]. Furthermore, DSGE models often assume rational expectations, i.e., agents know the constraints and dynamics of the entire economy, which has been criticized as philosophically unsound and empirically unjustified [13]. But, due to these assumptions, most DSGEs allow for a thorough analytical analysis.

ABMs allow implementing various individual decision models that are behaviorally more realistic than full economic rationality. Agents are often assumed to be boundedly rational and adapt their expectations, which is compatible with the Lucas critique [14]. In ABMs, fluctuations in aggregate variables arise not only from exogenous shocks as in DSGE models but primarily from irregularities in local interactions. Therefore, they offer an avenue for explaining various emergent phenomena [15] studied in empirical macroeconomics.

[*] kolb@pik-potsdam.de

On the other hand, ABMs are often very detailed so that an analytic treatment is unfeasible. Therefore, in ABMs, the difficulties arising from the aggregation of heterogeneous and interacting agents are usually solved computationally. Because the model mechanisms are difficult to trace in the 'black box' of a computational model, the results of ABMs are often difficult to interpret and cannot provide mathematically sound proofs of relationships between model variables. Results may therefore be difficult to generalize [16]. There has been some progress in the standardization of model descriptions for ABMs [17], but the lack of standardization, e.g., of decision rules, makes the models difficult to compare [5, p. 239]. Even though there are various techniques available for comprehensive model analysis [18], a systematic model exploration is uncommon and mostly limited to sensitivity analysis with respect to crucial parameters.

Methods from theoretical physics have been applied successfully to various problems in economics for many years [19]. Here, aggregation methods from statistical physics can bridge the gap between analytic macroeconomic models such as DSGE approaches and agent-based computational models (for a review of physics methods in social modeling, see Refs. [20] and [21]). In contrast to macroeconomic models, these approaches account for local interactions and use aggregation techniques to derive macrodynamics, providing a true microfoundation of the resulting macromodel. These kinds of approximation methods have found much interest in the fields of financial economics, behavioral finance, and evolutionary game theory recently and have produced interesting and promising results, e.g., to explain macroeconomic fluctuations (e.g., [22]) and understand propagation of financial shocks and the resulting systemic risk (e.g., [23]).

Many authors use mean-field approximations to aggregate interactions between heterogeneous agents, e.g., making use of stochastic differential equations or master or Fokker-Planck equations [24–33]. Such approaches assume that each agent pair interacts with the same probability. But many social and economic interactions are structured and the structure can be described by complex networks [34]. To also capture the dynamics arising from structured interactions, so-called moment closure methods take the microstructure of networks into account when deriving macroscopic quantities (e.g., [35], [36]). Thereby, they are able to show that often the network structure, whether fixed or evolving, has a crucial influence on the dynamics not only quantitatively but also qualitatively in enriching the stability landscape and introducing additional (meta-)stable dynamical regimes, e.g., due to effects related to clustering and community structure.

Yet, most of the literature regards either the network between agents or the states of agents as static, implicitly assuming different time scales for dynamics of and processes in the network. However, recent literature on opinion formation processes and the spreading of social norms in the field of computational social sciences suggests that both happen on a comparable time scale and therefore cannot be treated separately [7,37]. For such adaptive networks [7], moment closure techniques have been introduced in the physics literature to aggregate the feedback between complex adaptive network dynamics and dynamics of single-node states [38–41]. Here, we introduce these techniques to economic modeling and combine them with approaches from macroeconomics where interactions also happen globally via aggregated variables.

The technical challenges of analytic approximation methods for agent-based models has so far hampered their widespread use in economics. But they have a huge potential in providing profound insights into dynamical properties of economic systems: First, they help to increase the performance of computer simulations, making calculation of single model runs much faster and therefore allowing for a wider range of bifurcation and parameter analyses. Second, in contrast to stochastic simulations, they make formal proofs of relations between macroscopic variables possible. Third, they allow the derivation of analytical expressions of relations between model variables from the dynamic equations, which is not possible from single simulation runs. This paper takes a step forward in showcasing how such methods can be used to combine interactions in complex adaptive networks with macroeconomic modeling. It is therefore a contribution to the integration of nonstandard behavioral assumptions into macroeconomic models.

The agent-based model we introduce as an illustration of these methods is designed to investigate low-carbon transitions in an economy in the context climate economics and features both local interactions on a network and system-level interaction through markets. We use an adaptive network approach for our model to demonstrate how the individual approximation techniques mentioned above may be combined. In our model, the network of interactions between agents as well as the spreading of strategies between agents in this interaction network happens on a comparable time scale. In particular, we combine the different approximation techniques mentioned above, namely, moment closure, pair approximation, and large-system-limit approximations to derive an aggregate description for the dynamics of our model (for an overview of the different techniques, see [42]). The model consists of heterogeneous households that interact and learn from neighbors in a social network and a two-sector productive economy. The households differ in their investment strategies: they invest their savings either in the "dirty" or in the "clean" sector, each representing a separate capital market through which the agents interact. Agents imitate the investment strategy of acquaintances that are better off with a higher probability. To the best of our knowledge this is the first study that applies such a combination of approximation methods in a model that combines structured local with global interactions of heterogeneous agents in a socioeconomic setting. By successfully applying approximation techniques for adaptive networks to our model, we demonstrate that they are useful for investigating economic relationships within considerably complex models. Even though our reference application is an economic one, this approximation method can also be used to describe similarly structured models in other fields of research such as social ecology, neuroscience, and computational social science.

In the remainder of the paper, we first describe the details of the model (Sec. II). We then derive an aggregate description of the model by applying three approximation techniques: moment closure, pair approximation, and large-system limit (Sec. III). We discuss commonalities and differences between computer simulations and the approximation

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

approach. Before concluding, we illustrate how the derived macroapproximation can be used in a bifurcation analysis to better understand the qualitative properties of the nonlinear model (Sec. IV).

II. MODEL DESCRIPTION

To illustrate the use of the methods that we put forward, we develop a model of a stylized economy that captures the shift from a fossil-fuel-based to a renewable-energy-based sector. Decarbonization pathways consistent with the Paris agreement require a rapid shift of investments away from fossil fuel exploration and extraction to the development and deployment of renewable energies [43]. However, the implementation of climate policies is uncertain and expectations cannot be based on self-consistent beliefs about the future. In conventional macroeconomic models such shifts can only occur due to price signals from either improvements in green technology, increasing scarcity of fossil reserves, or carbon pricing. While price signals are certainly important, movements advocating for the divestment from fossil fuels point to the role of social norms and practices regarding investment decision to initiate and accelerate the energy transition [44]. To better understand such culturally driven situations of socioeconomic change, it is important to develop models that can incorporate endogenous preferences [45,46] and aspects of bounded rationality [47] such as imperfect foresight and information as well as learning.

Our model is designed to incorporate social dynamics that influence investment decisions [48,49]. In the context of climate economics and policy, the literature on social influence and norms has pointed out that such mechanisms are a leverage point to induce rapid change in socioeconomic systems [50–54]. The model focuses on two important mechanisms: First, investment strategies are spread on a network, which can be understood as a social learning process [55] influenced by social norms [56]. Second, the network adapts endogenously based on simple rules that model homophyly [57,58]. In the following, we explain the different parts of our two-sector model in detail.

A. Economic production

Our model as outlined in Fig. 1 consists of two sectors for production and a set of heterogeneous households that interact via a complex adaptive social network. The two production sectors employ different technologies. The production technology in one sector depends on the input of an exhaustible (fossil) energy resource R that is used up in the process, whereas the technology in the other sector does not. We call them the *dirty* and the *clean* sectors accordingly. We assume that physical capital is technology specific and cannot be reallocated between the two sectors. Therefore, the heterogeneous households in the model provide different types of capital K_j as well as labor L to the sectors. We assume that the technology in the dirty sector is fully developed and adequately described in terms of a fixed technological factor subsumed in the constant b_d, the so-called total factor productivity. For fossil fuels, price elasticities of demand, i.e., changes in demand in response to increasing or decreasing prices, are low in real economies [59–61], even with the choice between alternative technologies factored in. We approximate this by assuming that the fossil resource cannot be substituted by other production factors (capital, labor) in the dirty sector. This is in line with critique of the commonly assumed substitutability of natural resources in some widely used production functions in neoclassical models [62–66]. However, we acknowledge that a shift in the output of economic production from manufacturing to services can lead to substitution of resources by capital and labor [67] and argue that our model pictures this in a shift of economic production from the dirty to the clean sector, which is described in the following.

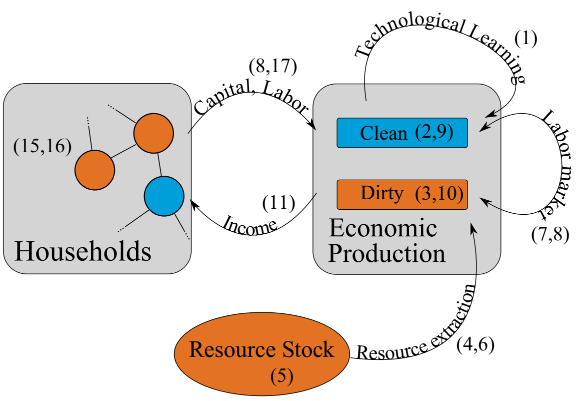


FIG. 1. Schematic of the model consisting of two production sectors of which one depends on an exhaustible fossil resource stock as well as a set of heterogeneous households that interact on an adaptive complex network and use social learning to decide upon which of two production sectors to invest in. Boxes and bubbles denote modeled entities; arrows denote interactions. Numbers in parentheses refer to equations that describe the specific part of the model.

The clean sector represents a circular economy in which the output of final goods depends on the machinery, knowledge, and effort used in its production and is not limited by resource scarcity on the time scale under consideration. The technology C used in the clean sector is assumed to be still in development and is therefore explicitly modeled. Following [68], we model technological progress as learning by doing according to Wright's law [69,70]. We assume that C is proportional to cumulative production but also depreciates with a constant rate χ. Depreciation can be regarded as a human capital effect that leads to knowledge depreciation over time as in [71]. This is also in line with the empirically observed decrease in learning rates for maturing technologies [68]

$$\dot{C} = Y_c - \chi C. \tag{1}$$

Capital, labor, and technology or knowledge are assumed to be mutual substitutes. To satisfy these requirements, we use the following production functions:

$$Y_c = b_c C^{\gamma} L_c^{\alpha_c} K_c^{\beta_c}, \tag{2}$$

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

$$Y_d = \min\left(b_d L_d^{\alpha_d} K_d^{\beta_d}, eR\right). \tag{3}$$

Subscripts c and d denote the clean and dirty sectors, respectively, L_c and L_d are labor in the two sectors, α and β are the elasticities of the respective input factors, b_c and b_d are the total factor productivities, and K_c and K_d are the capital stocks for the respective sector. Measuring unit production cost in the number of working hours as in the original study [69], γ is equivalent to the elasticity of learning by doing in the clean sector as outlined in [71].

We assume an efficient usage of resources in the dirty sector, such that

$$b_d L_d^{\alpha_d} K_d^{\beta_d} = eR, \tag{4}$$

where $1/e$ is the resource intensity of the sector, i.e., the amount of fossil resource needed for 1 unit of final product. The usage of the fossil resource R depletes a geological resource stock G with the initial stock $G(t=0) = G_0$:

$$\dot{G} = -R. \tag{5}$$

In line with the assumptions common in the literature [72,73], the cost of the fossil resource extraction and provision c_R depends on the resource flow R and the remaining fossil resource stock G such that $\partial c_R/\partial R > 0$ and $\partial c_R/\partial G < 0$. We chose the specific form to be

$$c_R = b_R R^{\rho} \left(\frac{G_0}{G}\right)^{\mu}, \quad \rho \geqslant 1, \quad \mu > 0, \tag{6}$$

such that at some point $\partial Y_d/\partial R < \partial c_R/\partial R$ to take into account that some part of the resource is not economic, i.e., its marginal cost exceeds its marginal productivity. We assume perfect labor mobility and competition for labor between the two sectors. This leads to an equilibrium wage w that equals the marginal return for labor, i.e., the production increase from an additional unit of labor,

$$w = \frac{\partial Y_c}{\partial L_c} = \frac{\partial Y_d}{\partial L_d} - \frac{\partial c_R}{\partial L_d}, \tag{7}$$

with the sum of labor in both sectors equal to a constant total amount of labor:

$$L_c + L_d = L. \tag{8}$$

As discussed before, we assume physical capital to be specific to the technology employed such that it can only be used in the sector in which it has been invested originally. This means that there are separate capital markets for the two sectors. We assume these capital markets to be fully competitive, resulting in capital rents equal to marginal productivity, after accounting for energy costs:

$$r_c = \frac{\partial Y_c}{\partial K_c}, \tag{9}$$

$$r_d = \frac{\partial Y_d}{\partial K_d} - \frac{\partial c_R}{\partial K_d}. \tag{10}$$

B. Adaptive network model for investment decision making

We model households as boundedly rational decision makers [74–76]: Households take their investment decisions, i.e., whether to invest their savings in the clean or the dirty sector, not by forming rational expectations [13,14] but by engaging in social learning [55] to obtain successful strategies [77] with reasonable effort. The outcomes of social learning crucially depend on the structural properties of the complex network of social ties among the households [78]. The strong and still increasing polarization of some societies on climate change issues suggests that social dynamics reinforce opposed positions in the population [79–84]. In static network models, such effects cannot be represented. Therefore, we model the adaptive formation of the social network endogenously. A well-established principle for the emergence of structured ties in social networks is homophily, i.e., the tendency that similar individuals get linked [57,85,86]. The following model specification uses social learning in combination with endogenous network formation based on homophily to model the investment decisions of the households.

We model N heterogeneous households denoted with the index i as owners of one unit of labor $L^{(i)} = L/N$ and capital $K_c^{(i)}$ and $K_d^{(i)}$ in the clean and dirty economic sectors, respectively. Households generate an income $I^{(i)}$ from their labor and capital income which they use for consumption $F^{(i)}$ and savings $S^{(i)}$. The rate at which households save their income is assumed to be fixed and is given by the savings rate s:

$$I^{(i)} = wL^{(i)} + r_c K_c^{(i)} + r_d K_d^{(i)}, \tag{11}$$

$$F^{(i)} = (1-s)I^{(i)}, \tag{12}$$

$$S^{(i)} = sI^{(i)}. \tag{13}$$

A binary decision parameter $o_i \in [c, d]$ denotes the sector in which the households decide to invest. As motivated above, we model decision making that is driven by two processes: social learning via the imitation of successful strategies and homophyly towards individuals exhibiting the same behavior.

We describe households as the nodes in a graph of acquaintance relations that change according to the following rules.

(1) Households get active at a constant rate $1/\tau$.

(2) When a household i becomes active, it interacts with one of its acquaintances j chosen uniformly at random.

(3) If they follow the same strategy, i.e., they invest in the same sector, nothing happens.

(4) If they follow a different strategy, i.e., they invest in different sectors, one of two actions can happen:

(a) Homophilic network adaptation: With probability φ, the households end their relation and household i connects to another household k, that follows the same strategy.

(b) Imitation: With probability $1-\varphi$, household i engages in social learning, i.e., it imitates the strategy of household j with a probability p_{ji} that increases with their difference in income.

We follow previous results on human strategy updating in repeated interactions from [77] when we assume the imitation probability as a monotonously increasing sigmoidal function of the relative difference in consumption between both households:

$$p_{ji} = \left(1 + \exp\left(-\frac{a(F^{(i)} - F^{(j)})}{F^{(i)} + F^{(j)}}\right)\right)^{-1}. \tag{14}$$

TABLE I. List of model parameters with their default values. Note that the parameter values are set to mirror plausible values observed in real-world economies but are not the result of a detailed model estimation procedure.

Symbol	Value	Parameter description
N	200	Number of households
M	2000	Number of network links between the households
b_c	1	Total factor productivity in the clean sector
b_d	4	Total factor productivity in the dirty sector
b_R	0.1	Initial resource extraction cost
e	1	Resource conversion efficiency
κ	0.06	Capital depreciation rate
χ	0.1	Knowledge depreciation rate
γ	0.1	Elasticity of knowledge in the clean sector
α_c	0.5	Elasticity of labor in the clean sector
α_d	0.5	Elasticity of labor in the dirty sector
β_c	0.5	Elasticity of capital in the clean sector
β_d	0.5	Elasticity of capital in the dirty sector
φ	0.5	Fraction of rewiring events in opinion formation
$1/\tau$	1.	Rate of opinion formation events
ε	0.05	Fraction of noise events in opinion formation
G_0	1 000 000	Initial resource stock
L	100	Total labor
s	0.25	Savings rate
ρ	1	Exponent for resource flow in extraction cost
μ	2	Exponent for resource stock in extraction cost

As opposed to the absolute difference in the original study [77], the probability in our model depends on relative differences. We set $a = 8$ to conform to their empirical evidence. This dependence on relative differences in per-household quantities is crucial for our method as we discuss at the end of Sec. III D. We model strategy exploration as a fraction ε of events that are random, e.g., rewiring to a random other household or randomly investing in one of the two sectors. Given the savings decisions of the individual households, and assuming equal capital depreciation rates κ in both sectors, the time development of their capital holdings is given by

$$\dot{K}_c^{(i)} = \delta_{o_i c} s\big(r_c K_c^{(i)} + r_d K_d^{(i)} + wL_i\big) - \kappa K_c^{(i)}, \tag{15}$$

$$\dot{K}_d^{(i)} = \delta_{o_i d} s\big(r_c K_c^{(i)} + r_d K_d^{(i)} + wL_i\big) - \kappa K_d^{(i)}, \tag{16}$$

where δ_{ij} is the Kronecker delta. The total capital stocks in the two sectors are made up of the sum of the individual capital stocks

$$K_j = \sum_i^N K_j^{(i)} = N k_j, \tag{17}$$

where k_j is the average per-household capital stock of a given capital type.

We acknowledge the fact that different model specifications are possible and interesting. For instance, we only consider fixed savings rates and the decision between two capital assets and leave the analysis of the interesting possible effects of households setting their savings rates individually to another study [87]. However, we want to point out that the approximation methods that we develop in the following are highly useful to gain insights from different but similar models that rely on complex adaptive interaction networks.

C. Numerical modelling and results

With the model specifications from Sec. II, the parametrization in Table I, and appropriate initial conditions for the dynamic variables, the model can be simulated numerically. For this, we implemented the dynamics in the multipurpose programming language PYTHON. The implementation of the ABM as well as the numerical analysis using the approximation methods described in the following is available at the github software versioning service in [88]. In the following, we discuss the resulting aggregate dynamics.

Figure 2 displays an exemplary average evolution of our model calculated as the mean of 100 simulation runs. The simulation starts with initial conditions of abundant fossil resources G and low clean technology knowledge stock C [Fig. 2(b)] as well as equally low capital stocks in the clean and dirty sectors K_c and K_d [Fig. 2(c)]. As we show later (see Sec. IV), the rest of the initial configuration of the model is rather irrelevant for the selected parameter values listed in Table I, since there is only one stable dynamical equilibrium as long as resource extraction costs are negligibly low. The high initial capital rents r_c and r_d are a direct result of our model assumptions and initial conditions, more precisely, the assumption that capital rent equals marginal productivity in Eqs. (9) and (10) and that of decreasing marginal productivity due to our choice of β_i in combination with the initial condition of low capital and a fixed labor supply. Also as a direct consequence of these assumptions, the capital rents r_c and r_d decrease over time as the capital stock is built up. Initially (from $t = 0$ to $t = 100$), as a result of our choice of total factor productivities b_i and due to low fossil resource extraction costs, capital productivity (and therefore capital rent r) is higher in the dirty sector than the clean sector [see Fig. 2(a)]. Consequently, the majority of households invest in the dirty sector, which leads to a high capital stock K_d [Fig. 2(c)] and high production output Y_d [Fig. 2(d)] in this sector.

Regarding the capital rents, we would expect the system to move towards a dynamic equilibrium in which the capital rent is equal in both sectors, i.e., $r_d = r_c$, if everything else remained constant. However, we find that there is a persisting difference between r_c and r_d between $t = 50$ and $t = 100$. This difference can be explained by the exploration of investment strategies even if they perform worse, which brings the shares of clean and dirty investors closer together. In terms of the depicted variables this means that it brings n_c closer to 0.5.

For $t > 100$ the depletion of the fossil resource leads to significantly increasing resource extraction costs. Consequently, the marginal productivity of dirty capital K_d decreases and so does r_d, leading to a peak in accumulation of capital in the dirty sector around $t = 100$ [Fig. 2(c)]. Once the relative return on capital in the clean sector increases, households start to adopt a clean investment strategy visible in an increase in n_c in Fig. 2(a). When the fossil resource stock reaches its economically exploitable share at around $t = 200$, the overall productivity in the dirty sector reaches 0, leading to full employment of all available labor in the clean sector. This drives demand for capital in the clean sector up, accelerating

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

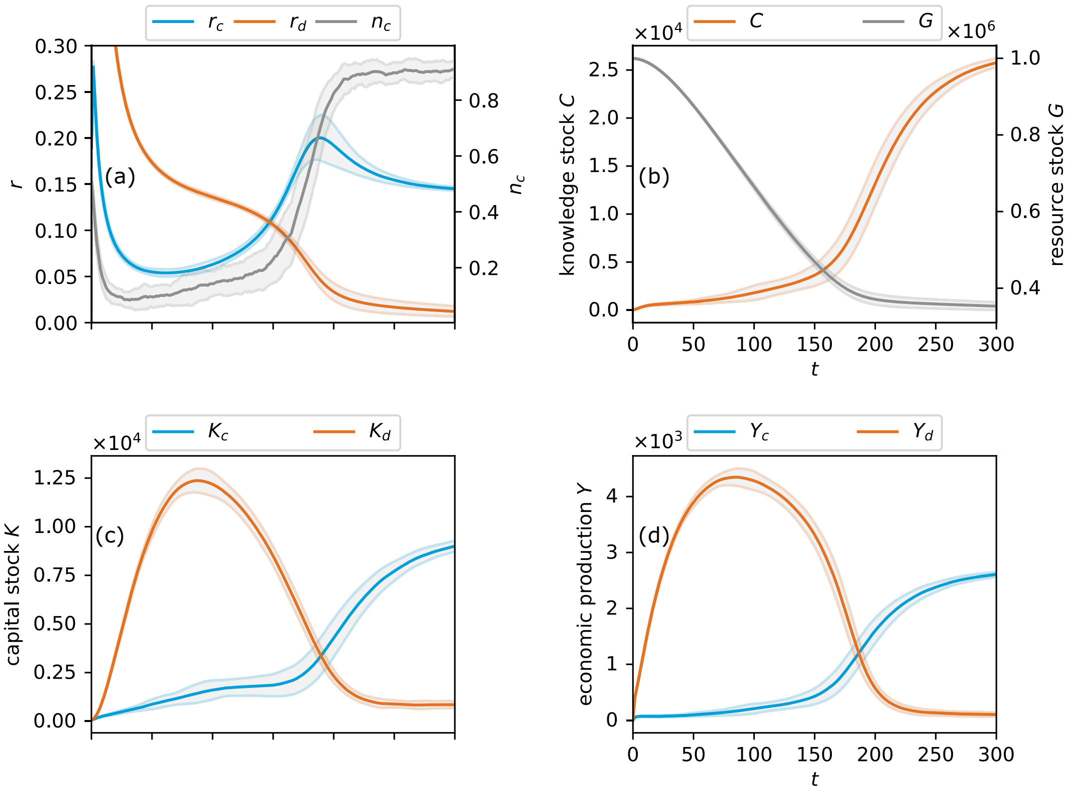


FIG. 2. Example trajectory of the ABM. Solid lines represent mean results from 100 runs of the model. Gray areas around solid lines show their standard deviation. Panels show capital rents in the clean and dirty sectors r_c and r_d as well as the fraction of households investing in the clean sector n_c in (a), knowledge and resource stock C and G in (b), output of the clean and dirty sectors Y_c and Y_d in (c), and capital stocks K_c and K_d in the clean and dirty sector in (d). Initial conditions are $G = G_0$, $C = 1$, $K_j^{(i)} = 1$ for the economic subsystem. For the investment decision process, the initial opinions of the $N = 200$ households are drawn from a uniform distribution. Their initial acquaintance structure is an Erdős-Renyi random graph with mean degree $k = 10$.

the change from dirty to clean investment. As all households except for the share caused by exploration are investing in the clean sector, the system reaches an equilibrium with high capital in the clean sector and low capital in the dirty sector.

Notably, we find an increasing variance in the fraction of households investing in the clean sector before and around the transition, which means that due to the stochasticity of the social learning process the transition happens earlier for some simulation runs than for others. Nevertheless, we find that the inertia of the model resulting from the large accumulated stock of capital that is specific to the dirty sector eventually leads to an almost-complete depletion of the fossil resource.

The adaptation dynamics in our model can lead to a fragmentation of the network with stark economic consequences. As the results in Appendix B show, an increased rewiring rate φ in the network adaptation process leads to a strongly delayed shift of investment from one sector to the other during the transition, even though the incentive in terms of an increased return r_c for the investment in this sector is high. This fragmentation is equivalent to a strong decline in the fraction of active edges in the network, e.g., the fraction of edges that connect households investing in different sectors of the economy. This finding is consistent with a major result of adaptive network modeling studies that show that adaptation will lead to fragmentation of a network at high rewiring rates φ [26,29,77,78,89]. Such network properties emerging from adaptation dynamics have been studied, for example, in the context of opinion dynamics, epidemics, and social-ecological systems [7,40,91,92]. One could suspect that the slowdown in the transition from one sector to the other results from the decreased rate of imitation events as their frequency scales with $1 - \varphi$. However, the results in Appendix A show that this effect is particular to the adaptive network model and cannot be reproduced in a well-mixed system simply by adjusting for the reduced frequency of imitation events. Appendixes B and A discuss further differences between the full model and special cases without adaptation as well as well-mixed interaction.

III. APPROXIMATE ANALYTICAL SOLUTION

Structurally, the model described in Sec. II consists of a set of coupled ordinary differential equations, (1), (5), (15), and (16), with algebraic constraints (4), (7), (8), (9) and (10) for the economic production process and a stochastic adaptive network process for the social learning component that is

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

described by rules 1 to 4 in Sec. II B. The state space of this combined process consists of 2 degrees of freedom of the knowledge stock and the geological resource stock as well as $2N$ degrees of freedom for the capital holdings of the set of all individual households plus the configuration space of the adaptive network process of the social learning component. We denote the variables of this process by capital letters $(C, G, K_j^{(i)}, \dots)$. To find an analytic description of the model in terms of a low-dimensional system of ordinary differential equations, we approximate it via a pair-based proxy process, a stochastic process in terms of aggregated quantities, thereby drastically reducing the dimensionality of the state space. We denote the variables of this process by capital letters with overbars $(\bar{X}, \bar{Y}, \bar{Z}, \bar{K}_l^{(k)}, \dots)$.

The derivation of this approximate process is done in three steps: First, we solve the algebraic constraints to the economic production process given by market clearing in the labor market and efficient production in the dirty sector—loosely following [93]. Second, we use a pair approximation to describe the complex adaptive network process of social learning in terms of aggregated variables, similarly to [91]. Third, we use a moment-closure method to approximate higher moments of the distribution of the capital holdings of the heterogeneous households by quantities related to the first moments of their distribution. Finally, we take the limit of infinitely many households (large-system or thermodynamic limit) to obtain a deterministic description of the system.

A. Algebraic constraints

To calculate labor L_c and L_d as well as wages in the two sectors, we use Eqs. (6) and (7) and for simplicity assume $\rho = 1$ and $\mu = 2$. We also assume equal labor elasticities in both sectors $\alpha_d = \alpha_c = \alpha$, resulting in

$$\begin{aligned} w &= \frac{\partial Y_d}{\partial L_d} - \frac{\partial c_R}{\partial L_d} \\ &= \frac{\partial Y_d}{\partial L_d} - \frac{\partial c_R}{\partial R}\frac{\partial R}{\partial L_d} = \frac{\partial Y_d}{\partial L_d} - \frac{\partial c_R}{\partial R}\frac{\partial}{\partial L_d}\frac{Y_d}{e} \\ &= \frac{\partial Y_d}{\partial L_d} - b_R\frac{G_0^2}{G^2}\frac{\partial}{\partial L_d}\frac{Y_d}{e} = b_d\alpha L_d^{\alpha-1}K_d^{\beta_d}\left(1 - \frac{b_R}{e}\frac{G_0^2}{G^2}\right) \end{aligned} \tag{18}$$

for the dirty sector and

$$w = b_c\alpha L_c^{\alpha-1}K_c^{\beta_c}C^\gamma \tag{19}$$

for the clean sector. Combining these results via Eq. (8), substituting

$$\begin{aligned} X_c &= \left(b_c K_c^{\beta_c} C^\gamma\right)^{\frac{1}{1-\alpha}}, \quad X_d = \left(b_d K_d^{\beta_d}\right)^{\frac{1}{1-\alpha}}, \\ X_R &= \left(1 - \frac{b_R}{e}\frac{G_0^2}{G^2}\right)^{\frac{1}{1-\alpha}}, \end{aligned} \tag{20}$$

and solving for w yields

$$w = \alpha L^{\alpha-1}(X_c + X_d X_R)^{1-\alpha}. \tag{21}$$

Plugging (21) into Eqs. (18) and (19) results in

$$L_c = L\frac{X_c}{X_c + X_d X_R}, \tag{22}$$

$$L_d = L\frac{X_d X_R}{X_c + X_d X_R} \tag{23}$$

for labor in the two sectors, and plugging this into (4) leads to

$$R = \frac{b_d}{e}K_d^{\beta_d}L^\alpha\left(\frac{X_d X_R}{X_c + X_d X_R}\right)^\alpha \tag{24}$$

for the use of the fossil resource. Using the results for L_c and L_d together with Eqs. (9) and (10), the return rates on capital result in

$$r_c = \frac{\beta_c}{K_c}X_c L^\alpha (X_c + X_d X_R)^{-\alpha}, \tag{25}$$

$$r_d = \frac{\beta_d}{K_d}(X_d X_R)L^\alpha (X_c + X_d X_R)^{-\alpha}. \tag{26}$$

It is also noteworthy that if we assume constant returns to scale with respect to capital and labor, e.g.,

$$\beta_c = \beta_d = 1 - \alpha \tag{27}$$

(even though it is not necessary for our method), this yields zero profits in both sectors:

$$\begin{aligned} Y_c &= wL_c + r_c K_c, \\ Y_d &= wL_d + r_d K_d + c_R. \end{aligned}$$

To sum up, we solved the algebraic constraints to the ordinary differential equations describing the economic production process resulting in the following equations:

$$\begin{aligned} X_c &= \left(b_c K_c^{\beta_c} C^\gamma\right)^{\frac{1}{1-\alpha}}, \quad X_d = \left(b_d K_d^{\beta_d}\right)^{\frac{1}{1-\alpha}}, \\ X_R &= \left(1 - \frac{b_R}{e}\frac{G_0^2}{G^2}\right)^{\frac{1}{1-\alpha}}, && \text{(28a)} \\ w &= \alpha L^{\alpha-1}(X_c + X_d X_R)^{1-\alpha}, && \text{(28b)} \\ r_c &= \frac{\beta_c}{K_c}X_c L^\alpha (X_c + X_d X_R)^{-\alpha}, && \text{(28c)} \\ r_d &= \frac{\beta_d}{K_d}X_d X_R L^\alpha (X_c + X_d X_R)^{-\alpha}, && \text{(28d)} \\ R &= \frac{b_d}{e}K_d^{\beta_d}L^\alpha\left(\frac{X_d X_R}{X_c + X_d X_R}\right)^\alpha, && \text{(28e)} \\ \dot{G} &= -R, && \text{(28f)} \\ \dot{K}_c^{(i)} &= s\delta_{o_i,c}\left(r_c K_c^{(i)} + r_d K_d^{(i)} + wL^{(i)}\right) - \kappa K_c^{(i)}, && \text{(28g)} \\ \dot{K}_d^{(i)} &= s\delta_{o_i,d}\left(r_c K_c^{(i)} + r_d K_d^{(i)} + wL^{(i)}\right) - \kappa K_d^{(i)}, && \text{(28h)} \\ \dot{C} &= Y_c - \chi C. && \text{(28i)} \end{aligned}$$

B. Pair approximation

To derive a macroscopic approximation of the social learning process described by rules 1 to 4 in Sec. II B, we make use of a pair-based proxy process that is derived via pair approximation from the adaptive network process. This proxy process is not equivalent but sufficiently close to the microscopic process approximating it in terms of aggregated quantities by making certain assumptions about the properties of their microscopic structure. The aggregated quantities of interest are the number of households investing in clean capital $N^{(c)}$, the number of households investing in dirty capital $N^{(d)}$, and

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

the number of links between agents in the same group, $[cc]$ and $[dd]$, as well as between the two groups, $[cd]$. Since the total number of households N and links M are fixed, these five variables reduce to 3 degrees of freedom, which we parametrize as follows:

$$\bar{X} = N^{(c)} - N^{(d)}, \quad \bar{Y} = [cc] - [dd], \quad \bar{Z} = [cd]. \tag{29}$$

These 3 degrees of freedom span the reduced state space of the social process $\bar{\mathbf{S}} = (\bar{X}, \bar{Y}, \bar{Z})^T$. The investment decision-making process can then be described in terms of jump lengths $\Delta\bar{\mathbf{S}}_j$ and jump rates $W(\bar{\mathbf{S}}, \bar{\mathbf{S}} + \Delta\bar{\mathbf{S}}_j)$ in this state space for the different events j in the set Ω of all possible events. Their derivation is illustrated by the example of a clean household imitating a dirty household: The approximate rate of this event is given by

$$W_{c\to d} = \frac{N}{\tau}(1-\varepsilon)(1-\varphi)\frac{N^{(c)}}{N}\frac{[cd]}{[cd]+2[cc]}p_{cd}. \tag{30}$$

In some more detail this results from

(i) N/τ, the rate of social update events, i.e., the rate of events per household times the number of households.

(ii) $(1-\varepsilon)$, the probability of the event not being a noise event.

(iii) $(1-\varphi)$, the probability of imitation events (versus network adaptation events).

(iv) $N^{(c)}/N$, the probability that each active household will invest in clean capital.

(v) $[cd]/(2[cc]+[cd])$, the approximate probability of interaction with a household investing in dirty capital. Here, we approximate the distribution of dirty neighbors among clean households with its first moment i.e., we act as if links between clean and dirty households were evenly distributed among all households.

(vi) p_{cd}, the expected value of the probability that each active household will imitate its randomly chosen neighbor, depending on the difference in consumption between households investing in clean vs dirty capital as given in Eq. (14). The expression is derived in detail as part of the moment closure in Sec. III C.

The corresponding change in the state-space variables is a little trickier. Since the event is a clean household imitating a dirty household, we already know about one of the neighbors of the household. As laid out in detail in, e.g., [38], the state of the remaining neighbors in the full model is determined by the frequency of higher-order network motifs, e.g., $[dcd]$ and $[dcc]$. The frequency of these higher-order motifs is approximated by the expected value of the states of additional neighbors as follows: summing over the excess degree of node q^c by drawing $k^c - 1$ times from the distribution of neighbors, which is, as before, approximated by an even distribution of edges between same and different households among all households. Again, this approximates the respective full distributions with their first moments. If one wanted to include higher-order effects in the network dynamics, one could follow one of the various ways laid out in, e.g., [39]. Thus the probability that a neighbor is dirty, $p^{(d)}$, or clean, $p^{(c)}$, reads

$$p^{(c)} = \frac{2[cc]}{2[cc]+[cd]}; \quad p^{(d)} = \frac{[cd]}{2[cc]+[cd]}. \tag{31}$$

This results in an expected number of $n^{(c)}$ additional clean neighbors and $n^{(d)}$ additional dirty neighbors,

$$n^{(c)} = (1 - 1/k^{(c)})\frac{2[cc]}{N^{(c)}}, \quad n^{(d)} = (1 - 1/k^{(c)})\frac{[cd]}{N^{(c)}}, \tag{32}$$

where $k^{(c)}$ is the mean degree, e.g., the mean number of neighbors of a clean household in the network. With the results from (32) the changes in the expected values of the state space variables can be approximated as follows:

$$\begin{aligned}
\Delta N^{(c)} &= -1,\\
\Delta N^{(d)} &= 1,\\
\Delta[cc] &\approx \left(1 - \frac{1}{k^{(c)}}\right)\frac{2[cc]}{N^{(c)}},\\
\Delta[dd] &\approx \left(1 - \frac{1}{k^{(c)}}\right)\frac{[cd]}{N^{(c)}},\\
\Delta[cd] &\approx -1 + \left(1 - \frac{1}{k^{(c)}}\right)\frac{2[cc]-[cd]}{N^{(c)}},
\end{aligned}$$

and summing up, the change in the state vector is approximately given by

$$\Delta\bar{\mathbf{S}}_{c\to d} \approx \begin{pmatrix} -2 \\ -k^{(c)} \\ -1 + \left(1 - \frac{1}{k^{(c)}}\right)\frac{2[cc]-[cd]}{N^{(c)}} \end{pmatrix}. \tag{33}$$

In terms of the jump lengths $\Delta\bar{\mathbf{S}}$ and the rates W, the dynamics of the pair-based proxy can be written as a master equation for the probability distribution P in the state space of $\bar{\mathbf{S}}$:

$$\begin{aligned}
\frac{\partial P(\bar{\mathbf{S}}, t)}{\partial t} = \sum_{j\in\Omega} & P(\bar{\mathbf{S}} - \Delta\bar{\mathbf{S}}_j, t)W(\bar{\mathbf{S}} - \Delta\bar{\mathbf{S}}_j, \bar{\mathbf{S}})\\
& - P(\bar{\mathbf{S}}, t)W(\bar{\mathbf{S}}, \bar{\mathbf{S}} + \Delta\bar{\mathbf{S}}_j).
\end{aligned} \tag{34}$$

C. Moment closure

To describe the capital structure in the model that consists of $2N$ equations of the type of (15) and (16), we use the cohort of $N^{(c)}$ households investing in clean and the cohort of $N^{(d)}$ households investing in dirty capital and look at the aggregates of their respective capital holdings:

$$\bar{K}_l^{(k)} = \sum_i^N \delta_{o_i k} K_l^{(i)}. \tag{35}$$

Here, the upper index in $\bar{K}_l^{(k)}$ indicates the shared investment decision of the cohort of households as opposed to the index of the individual household before. The lower index still denotes the capital type. $\delta_{o_i k}$ is the Kronecker delta.

Later, we use the fact that in the limit of $N \to \infty$ these aggregates should converge to their expected values, e.g., the first moments of their distribution with probability 1. The time derivative of the aggregates defined in (35) is given by the deterministic process of capital accumulation, (28g) and (28h), as well as terms resulting from the stochastic process

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

of agents switching their saving decisions:

$$\left.\begin{aligned}\dot{\bar{K}}_c^{(c)} &= (sr_c - \alpha)\bar{K}_c^{(c)} + sr_d\bar{K}_d^{(c)} + sw\bar{L}\\ \dot{\bar{K}}_d^{(c)} &= -\alpha\bar{K}_d^{(c)}\\ \dot{\bar{K}}_c^{(d)} &= -\alpha\bar{K}_c^{(d)}\\ \dot{\bar{K}}_d^{(d)} &= \underbrace{sr_c\bar{K}_c^{(d)} + (sr_d - \alpha)\bar{K}_d^{(d)} + sw\bar{L}}_{D_l^{(i)}}\end{aligned}\right. \quad +\text{switching terms.} \tag{36}$$

The switching terms for $\bar{K}_c^{(c)}$ result from agents changing their saving decision, thereby moving their capital endowments from the aggregate capital of the cohort of clean investors to the aggregate of the cohort of dirty investors, and vice versa. We assume that each household switching to the other cohort is endowed with the mean capital of the cohort and that their capital endowment is independent of the probability of switching such that we can describe the switching terms as a product of both factors. Then we can write down the changes in capital stocks explicitly including the switching terms as a simple stochastic differential equation,

$$\mathrm{d}\bar{K}_l^{(k)} = D_l^{(k)}\mathrm{d}t + \underbrace{\frac{\bar{K}_l^{(j)}}{N^{(j)}}\mathrm{d}N^{j\to k} - \frac{\bar{K}_l^{(k)}}{N^{(k)}}\mathrm{d}N^{k\to j}}_{\text{switching terms}}, \tag{37}$$

where the first term on the right-hand side refers to the change in aggregates without switching, as given by the equations of capital accumulation, (36), and the following terms denote the influx and outflux of capital from the aggregate due to households changing their savings decisions. $\mathrm{d}N^{j\to k}$ denotes the stochastic process of households switching from one opinion to another according to the rules outlined in Sec. II B. In line with the pair approximation described in Sec. III B we approximate them as

$$\mathrm{d}N^{j\to k} = \sum_{l\in\Omega_{j\to k}} W_l\mathrm{d}t, \tag{38}$$

where $\Omega_{j\to k}$ denotes the set of all events that result in a household changing from cohort j to cohort k and W_l is the rate of the respective event analogously to (30).

The imitation probability p_{cd} in Eq. (30) is approximated as the expected value of a linearized version of Eq. (14) when drawing a pair of neighboring households i, j as specified. More precisely, we perform a Taylor expansion of Eq. (14) in terms of the consumption of the two interacting households $F^{(c)}$ and $F^{(d)}$ around some fixed values $F^{(c)*}$ and $F^{(d)*}$ up to linear order. To maintain the symmetry of the imitation probabilities with respect to the household incomes, we change variables to $\Delta F = F^{(c)} - F^{(d)}$ and $F = F^{(c)} + F^{(d)}$ and expand around $\Delta F = 0$, $F = F_0$, where F_0 is yet to be fixed to a value. In linear order this results in

$$p_{cd} = \frac{1}{2} - \frac{a}{4F_0}\Delta F, \tag{39}$$

$$p_{dc} = \frac{1}{2} + \frac{a}{4F_0}\Delta F. \tag{40}$$

To make the approximation work in the biggest part of the system's state space, we set the reference point F_0 to be the middle of the sum of the estimated upper and lower bounds for the attainable income of households investing in the clean (dirty) sector. The minimum attainable income is assumed to be 0. The maximum attainable income for a household investing in the clean sector is assumed to be reached at equilibrium given that all other households also invest in the clean sector; e.g., we calculate $F^{(c)*}$ as half of an average household income at the steady state of $\dot{K}_c = sb_cL^\alpha K_c^{\beta_c}C^\gamma - \delta K_c$ and $\dot{C} = b_cL^\alpha K_c^{\beta_c}C^\gamma - \delta C$,

$$C^* = \left(\frac{b_cL^\alpha s^{\beta_c}}{\delta}\right)^{\frac{1}{1-\beta_c-\gamma}}, \quad K_c^* = \left(\frac{b_cL^\alpha s^{1-\gamma}}{\delta}\right)^{\frac{1}{1-\beta_c-\gamma}}. \tag{41}$$

Equivalently, we calculate $F^{(d)*}$ as half of an average household income at the steady state of $\dot{K}_d = s(1 - \frac{b_R}{e})b_dK_d^{\beta_d}P^\alpha - \delta K_d$:

$$K_d^* = \left(\frac{sb_dL^\alpha}{\delta}\left(1 - \frac{b_R}{e}\right)\right)^{\left(\frac{1}{1-\beta_d}\right)}. \tag{42}$$

With these results, using the fact that we set $\beta_c = \beta_d = \alpha = 1/2$, the reference point F_0 is

$$\begin{aligned} F_0 &= \frac{1}{2}\big(F^{(c)*} + F^{(d)*}\big)\\ &= \frac{1-s}{2N}(r_c^*K_c^* + wL + r_d^*K_d^* + wL) \qquad (43)\\ &= \frac{1-s}{2N}\left(\left(\frac{sb_cL^\alpha}{\delta^{\beta_c+\gamma}}\right)^{\frac{1}{1-\beta_c-\gamma}} + \frac{s}{\delta}\left(\left(1 - \frac{b_R}{e}\right)b_dL^\alpha\right)^2\right), \qquad (44)\end{aligned}$$

where r_c^* and r_d^* in (43) are the capital return rates, (9) and (10), in the respective equilibria, (41) and (42).

Given this linear approximation of the imitation probabilities, we approximate the consumption F_c and F_d of the randomly selected households i and j as the household consumption of the average household investing in clean and dirty capital using the aggregated variables as introduced in (35). In the large-system limit, this is equivalent to taking the expected value over all households in the respective cohorts:

$$\begin{aligned} p_{cd} = \frac{1}{2} - \frac{a}{4F_0}\Big(&r_c\big(\bar{K}_c^{(c)} - \bar{K}_c^{(d)}\big) + r_d\big(\bar{K}_d^{(c)} - \bar{K}_d^{(d)}\big)\\ &+ w\frac{L}{N}\big(N^{(c)} - N^{(d)}\big)\Big), \end{aligned} \tag{45}$$

$$\begin{aligned} p_{dc} = \frac{1}{2} + \frac{a}{4F_0}\Big(&r_c\big(\bar{K}_c^{(c)} - \bar{K}_c^{(d)}\big) + r_d\big(\bar{K}_d^{(c)} - \bar{K}_d^{(d)}\big)\\ &+ w\frac{L}{N}\big(N^{(c)} - N^{(d)}\big)\Big). \end{aligned} \tag{46}$$

With this approximation, we have now reached an approximate description of the microscopic dynamics in terms of stochastic differential equations for the aggregate variables.

D. Large-system limit

The description of the model in terms of Eqs. (28f), (28i) (34), and (36) poses a significant reduction in complexity, yet it is still a description in terms of a stochastic process rather than in terms of ordinary differential equations, as

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311.

typically used in macroeconomic models. To further reduce it to ordinary differential equations, we do an expansion in terms of system size, which in our case is given by the number of households N. Therefore, following Van Kampen [94, p. 244], we introduce the rescaled variables

$$x = \frac{X}{N}, \quad y = \frac{Y}{M}, \quad z = \frac{Z}{M}, \quad k = \frac{2M}{N} \tag{47}$$

and expand the master equation, (34), that describes the social learning process in terms of a small parameter N^{-1}. In the leading order, the time development of the rescaled state vector $\mathbf{s} = (x, y, z)$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{s} = \alpha_{1,0}(\mathbf{s}), \tag{48}$$

where $\alpha_{1,0}$ is the first jump moment of W. In terms of the rescaled variables $\mathbf{s}$, $\alpha_{1,0}$ is given by

$$\alpha_{1,0}(\mathbf{s}) = \int \Delta\mathbf{s} W(s, \Delta\mathbf{s})\mathrm{d}\Delta\mathbf{s}, \tag{49}$$

which in the case of discrete jumps in state space simplifies to

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{s} = \sum_{j\in\Omega} \Delta\mathbf{s}_{\mathbf{j}} W_j, \tag{50}$$

where Ω is the set of all possible (discrete) events in the opinion formation process.

As for the economic processes, we keep the aggregated quantities $(\bar{K}_i^j, C, G)$ fixed and formally go to a continuum of infinitesimally small households. As people and also households, for that matter, are finite entities, a continuum of households makes no sense. But practically, this can be understood as an interpretation of the heterogeneous households as a weighted sample of a very large population of heterogeneous individuals and increasing the sample size up to the point where a continuum of households is a sufficiently good approximation of reality in terms of the model. The only element in the approximation of the economic model that depends on per-household quantities is the imitation probability, (14), or rather its approximation, (39) and (40). Since we have chosen this to depend on relative differences in income, their dependence on the number of households N cancels out and the limit of $N \to \infty$ becomes trivial, resulting in the following deterministic approximation for the capital endowments in sector l of households investing in sector k described in Eq. (37),

$$\dot{\bar{K}}_l^{(k)} = D_l^{(k)} + \frac{\bar{K}_l^{(j)}}{N^{(j)}} \sum_{l\in\Omega_{j\to k}} W_l - \frac{\bar{K}_l^{(k)}}{N^{(k)}} \sum_{l\in\Omega_{k\to j}} W_l, \tag{51}$$

where $D_l^{(k)}$ are the capital accumulation terms as given in (36) and $\Omega_{l\to k}$ is the set of all opinion formation events, where a household changes its opinion from l to k.

Together with Eqs. (28f) and (28i) the sets of equations specified by (50) and (51) fully describe the approximate dynamics of the original model as specified in Sec. II. The full set of equations is given in Appendix C.

Our approximation reduces the full model to a set of first-order differential equations with 9 degrees of freedom. For comparison, the full model has $2N + 2$ degrees of freedom in the economic system plus the configuration space of the social network component. The right-hand sides of the set of differential equations are continuously differentiable and depend on 12 parameters for the economic system and 2 parameters for the social network process. The state space of the system is bounded between -1 and 1 in x and y and between 0 and 1 in z as well as by 0 from below in the variables of the economic system $\bar{K}_l^{(k)}$, G, and C. As the equations are bulky, it is recommended to use a computer algebra system to work with them.

The freedom to choose equations for economic production that are not scale invariant critically depends on the assumption that household interaction only depends on relative differences. For individual interaction that depends on absolute differences, one can show that the large-system limit only works if the system is scale invariant in terms of aggregated quantities. Nevertheless, it would be possible to relax both of these assumptions and to work with the pair-based proxy process with the results explicitly depending on the number of households, which in return could lead to interesting finite-size effects.

E. Results of the model approximation

The results in Fig. 3 are to some extent complementary to the results in Fig. 2 that we discussed in Sec. II C. Figure 3(d) shows capital in both sectors belonging to households that actually invest in these sectors, which is almost equivalent to the variables in Fig. 2(d), as it makes up almost the entirety of these capital stocks. This can be seen in Fig. 3(c): It shows the capital of households in the sector in which they do not currently invest, which is approximately an order of magnitude smaller (note the different scale of the vertical axis in the figure).

A comparison of the results of the approximation (dashed lines) with those of the numerical simulation of the ABM (solid lines) in Fig. 3 shows that the approximation exhibits the same qualitative features, such as the trends, timing, and order of magnitude of the displayed variables, as the microscopic model. Particularly, these results show that for the given parameter values the macroscopic approximation is capable of reproducing very closely the quasiequilibrium states before and after the transition from the dirty to the clean sector, as it lies within the standard error of the ensemble of ABM runs. Also, the approximation is reasonably capable of reproducing the timing of and the transient states during the transition. This is somewhat surprising since in other works, macroapproximations were less well able to get the timing of the transition right.

In the following, we discuss the existing differences between the results of the approximated model and the numerical simulation results. For instance, we find that the approximation estimates the transition from investment in the dirty sector to investment in the clean sector a bit too early [best visible in Fig. 3(a)]. The reason for this might be the slight underestimation of the share of clean-investing households, leading to a slight overestimation of the share of dirty capital in the system, which is also visible in Fig. 3(c).

We find a second obvious discrepancy between the micromodel and the approximation in the overestimation of

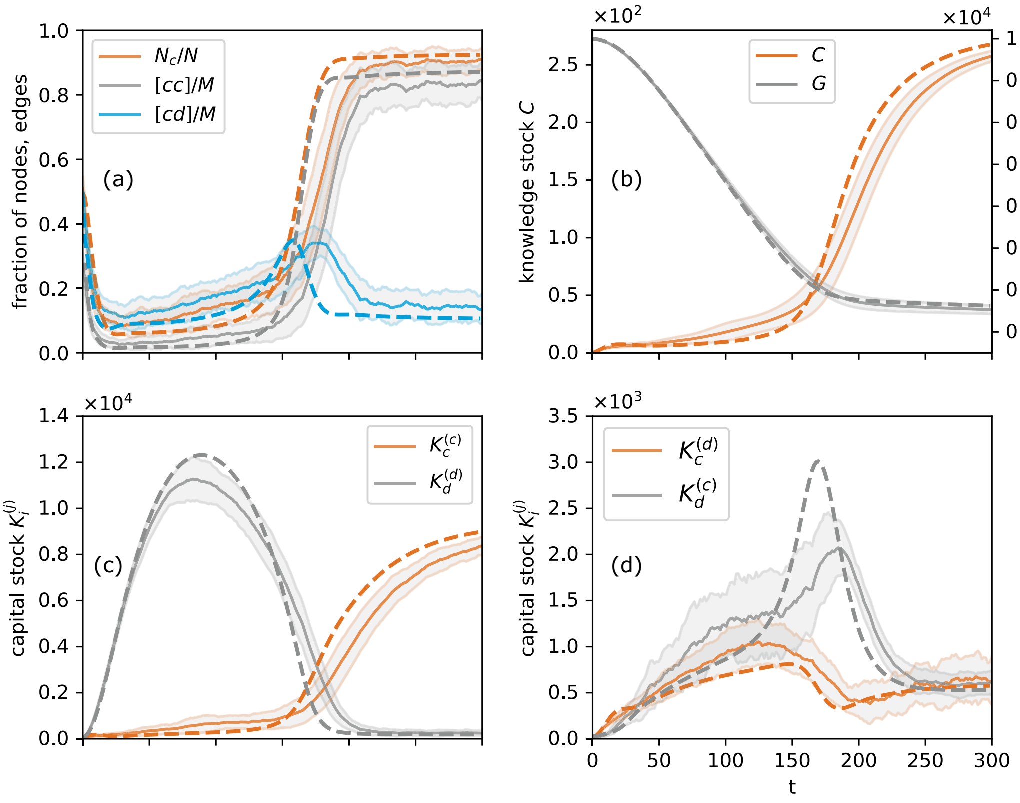


FIG. 3. Trajectories of dynamic variables from the macro approximation and from measurement in ABM simulations. The results of ABM simulations (solid lines) are obtained as an ensemble average of 50 runs, with standard errors indicated by gray areas. Initial conditions are given by equal shares of the $N = 200$ households investing in both sectors and equal endowments in both sectors for all households. The initial acquaintance network among the households is an Erdős-Renyi random graph with mean degree $k = 10$. Other initial conditions are $C_0 = 0.5$ and $G_0 = 5 \times 10^5$. All other parameters are listed in Table I. The results from the macro approximation (dashed lines of the same colors) are obtained by integration of the ODEs that are obtained from the large-system limit with fixed per-household quantities. The initial conditions are drawn from the same distribution as previously for the ABM simulations, e.g., N_c, $[cc]$, and $[cd]$ are calculated from an Erdős-Renyi random graph with mean degree $k = 10$.

dirty capital of clean investors ($K_d^{(c)}$) [Fig. 3(d)] during the transition phase between $t \approx 150$ and $t \approx 200$. This can be explained by the inequality in capital holdings among households. In the approximation, all households investing in dirty or clean capital are assumed to have the same income, respectively. Therefore, the probability of changing their investment behavior will change for all of them at once during the transition phase, leading to a rapid shift of dirty investors changing to invest in clean capital but taking their dirty capital endowments with them [hence the sharp peak in dirty capital of clean investors during the transition phase; see Fig. 3(d), upper dashed line].

Also, in the micromodel, households changing from a dirty to a clean investment strategy take their—presumably high—endowments in dirty capital with them. Therefore, the endowments in dirty capital of households investing in the clean sector are relatively widespread [see gray area around the upper solid line in Fig. 3(d)]. This has effects on the estimated timing of the transition too. In the micromodel, the income of households is heterogeneous. Therefore, for each of them the probability of changing their investment behavior changes at different points in time, i.e., poorer households are likely to switch earlier during the transition than richer households. Together this leads to a slower, more spread-out transition dynamic, the micromodel resulting in a flatter peak in the dirty capital endowments of clean-investing households.

Another effect at play during the transition is related to the assumptions in Eqs. (31) and (32). Namely, all households that invest in the same type of capital have the same distribution of clean and dirty neighbors.

In the reality of the micromodel, however, these assumptions that are essential to the pair approximation may well be wrong—especially so during a rapid transition. For example, a household that has only recently changed its state has a neighborhood that is atypical for its group and adapts only slowly. Consequently, when many changes in the state of the system happen in a short time, a significant proportion of the

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

population is not well described by the assumed approximate distribution.

A number of these effects that lead to discrepancies between the micromodel and the approximation can be mitigated by higher-order moment closure for the distribution of heterogeneous agent properties or higher-order motif approximation of the network dynamic.

For instance, a higher-order moment closure approximation that tracks the variance and skewness of the distribution of capital endowments can also account for the likelihood of capital endowments of agents that switch their investment decision to be biased. This would presumably mitigate the overestimation of dirty capital of clean investors $[K_d^{(c)}]$ during the transition as well as the underestimation of $[K_d^{(c)}]$ before the transition and therefore also estimate the timing of the transition even more precisely.

Similarly, a higher-order motif approximation of the network dynamic can describe the heterogeneity in the local distribution of opinions in the neighborhood of individual agents and correct for the effects of this, especially during periods of transient nonequilibrium dynamics in the approximated model.

In the previous section we derived a set of ordinary differential equations describing the stochastic dynamics of an agent-based model in terms of aggregated variables in the large-system limit. We intend this derivation to be a prototypical example for a macroeconomic model with true microfoundations based on heterogeneous agents, given that their microscopic interactions are of similar complexity. As such, it might also serve as a starting point for the application and development of similar models for other kinds of social dynamics. For example, an extension to continuous opinions requiring a Fokker-Planck-type description would follow naturally and would grant compatibility to a large body of models for social influence (see Ref. [95], pp. 988 ff.).

IV. BIFURCATION ANALYSIS

The description of the model as a system of ordinary differential equations allows for the analytical analysis of emergent model properties such as multistability, tipping, and phase transitions. As a proof of concept application we subsequently show the results of a bifurcation analysis.

A. Methods

Bifurcation theory is the analysis of qualitative changes of dynamical systems under parameter variation, for example, between a regime with a unique equilibrium (fixed point) and a multistable regime. The parameter value at which a qualitative change, for example, in the stability of an equilibrium, occurs is called a critical value or bifurcation point. Bifurcations are classified according to the changes in dynamical properties of the system [96,97]. Analytical methods have limited scope to identify bifurcation points in nonlinear systems. Methods like numerical continuation can handle complex systems of ordinary differential equations like the one derived in Sec. III [98]. Consequently, we use numerical continuation from PyDSTool [99,100], a PYTHON package for dynamical systems modeling and analysis [101].

A common bifurcation type that appears in our model is the fold bifurcation, which is also known as saddle-node bifurcation. This type is a local bifurcation in which a stable fixed point collides with an unstable one and both disappear.

Varying two bifurcation parameters at the same time can result in even richer qualitative changes in the dynamics. A prevalent example of such a bifurcation is the cusp geometry [97, p. 397]. A change in the second bifurcation parameter in this geometry beyond a certain value results in the so-called cusp catastrophe: the multistability of the system disappears for all values of the first bifurcation parameter. As we show in the following, the macroapproximation of our model indeed exhibits a cusp bifurcation.

B. Discussion of results

A considerable advantage of the description of our model in terms of ordinary differential equations (28f), (28i), (50), and (51) over agent-based modeling is the fact that it allows for the usage of established tools for bifurcation analysis. As a proof of concept, we show some results in Fig. 4. Here, we analyze the possible steady states of the system with abundant fossil resources, e.g., the possible equilibrium states of the model in the regime before the fossil resource becomes scarce and acts as an external driver on the system, pushing it towards clean investment. Therefore, we set the resource depletion to 0, i.e., we keep the resource stock in Eq. (28f) constant, $G(t) \equiv G_0$, such that the resource usage cost in Eq. (6) still depends on resource use R but is not increased by deceasing resource stock G. Thereby, we eliminate the rising resource extraction cost as the constraint in (7) and (10) that eventually halts production in the dirty sector. We choose the learning rate γ as the bifurcation parameter, as we expect it to yield interesting results. Generally, in nonlinear dynamical systems, exponential factors are expected to have a strong influence on dynamical properties. Therefore, changing these factors is expected to lead to bifurcation behavior. Consequently, in Figs. 4(a) and 4(c) we see that for certain learning rates γ the macroscopic approximation exhibits a bistable regime limited by two fold bifurcations with bifurcation points indicated by LP1 and LP2. In this regime both low investment in the clean sector together with high investment in the dirty sector and low knowledge as well as high investment in the clean sector together with low investment in the dirty sector and high knowledge are stable states of the economic system. This means that in this region economic outcomes are highly path dependent. Starting with slightly different knowledge about clean technologies may lead to widely differing adoption levels of the technology in the long run.

Figure 5 shows an example of how this bifurcation structure of the dynamical system depends on other parameters. When the total factor productivity in the dirty sector, b_d, is varied the system undergoes a cusp bifurcation. Above a certain value of b_d the system exhibits bistability, whereas below this value it does not.

Clearly, this choice of bifurcation parameters is only one of many, and other choices may very well lead to interesting results. However, we had to limit ourselves to this proof-of-concept study since an extensive analysis of all possible combinations would be well beyond the scope of this paper.

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

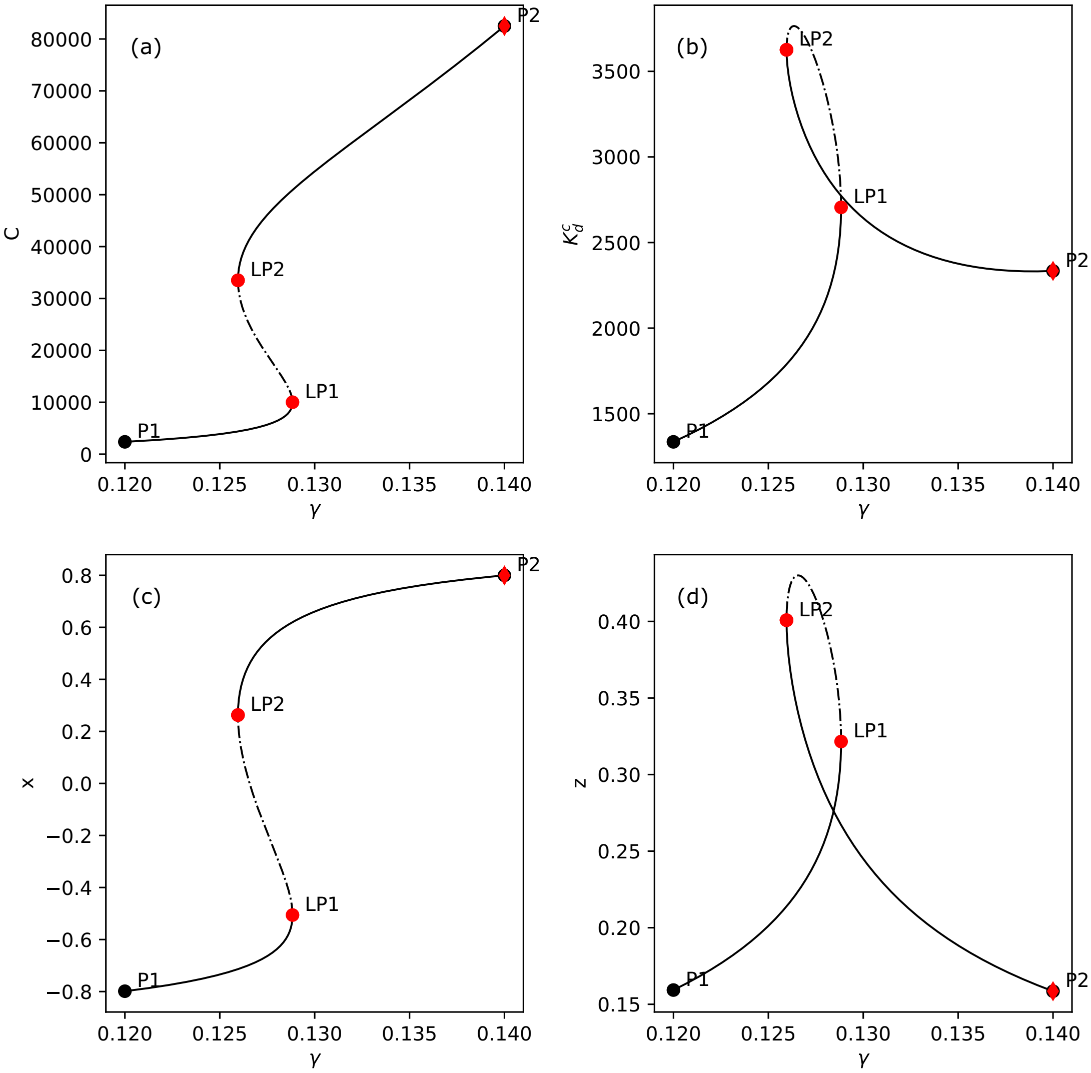


FIG. 4. Bifurcation diagram: Continuation of the stationary solution of the macroscopic approximation without resource depletion, i.e., with $\dot{G} = 0$ instead of the rate R as given by Eq. (28f). The bifurcation parameter is γ, the elasticity of knowledge in the clean sector, which also reflects the elasticity of learning by doing of the respective technology. The points labeled P1 and P2 are the beginning and end points of the continuation line; points LP1 and LP2 are the bifurcation points of twofold bifurcations. The stable unstable manifold is indicated by the dotted line; the stable manifold is indicated by the solid line. Note that the intersections of the curves in (b) and (d) do not actually mean that the stationary manifold is not a bijective function of the bifurcation parameter γ but rather a result of the projection of the multidimensional manifold onto the two-dimensional space.

Multistability of the economy would mean that policies could make use of inherent dynamical properties of the system to reach a desired state or bring the system onto a desired pathway. For example, policy measures such as regulation or taxes can help drive the system into another basin of attraction, i.e., a region of the phase space in which trajectories approach another equilibrium in the long term. To do so, the system has to cross a separatrix, the boundary between two basins of attraction. After this boundary is crossed, the policy measure can be discontinued, and the system's dynamics guarantee that it reaches the new equilibrium. Figure 5 shows that such an intervention could be complemented by an additional policy measure, lowering the total factor productivity in the dirty sector, effectively reducing the distance of the stable manifold from the separatrix and thereby presumably making the first measure less costly. Another possibility to take advantage of the system's inherent dynamical structure is to use its hysteresis, i.e., to find policy measures that change the first bifurcation parameter γ across a bifurcation point or to change the second bifurcation parameter b_d to move the bifurcation point past the current state of the system (or a combination of both), after which the system would fall to the other

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

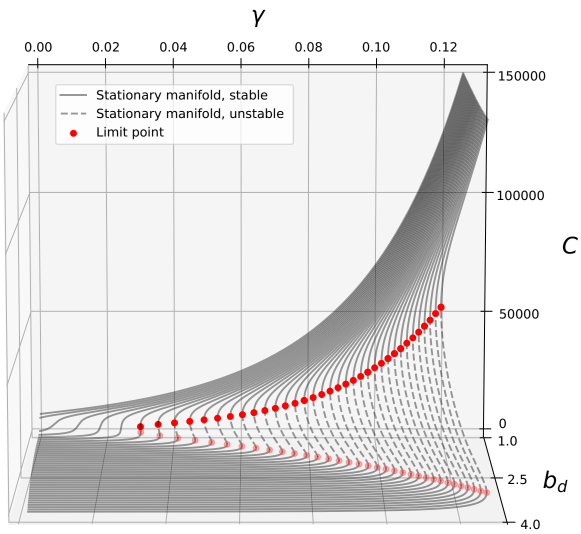


FIG. 5. Cusp bifurcation diagram: Stationary manifold from Fig. 4(a) for different values of the total factor productivity in the dirty sector b_d. Red circles indicate the limit points of the one-dimensional fold bifurcation separating the stable and the unstable parts of the stationary manifold indicated by the solid and the dashed line, respectively. For a critical value of $b_d \approx 1.4$ and $\gamma \approx 0.03034$ the two limit points converge and annihilate each other. This codimension 2 bifurcation with bifurcation parameters γ and b_d is called a cusp catastrophe. In our two-sector economic model, this results in a lock-in effect in the dirty sector; i.e., below this point, there is a smooth transition of production from the dirty to the clean sector and above this point production in the dirty sector is continued even though production in the clean sector would be more efficient.

branch of the stable manifold. Afterwards, the policy can be discontinued and the system would remain in its new state. For such considerations, tools from dynamical systems theory and topology can be used to classify the phase space of the system into regions with respect to the reachability of a desirable state [93,102]. This allows designing temporary policies that leverage the multistability of the socioeconomic system.

V. CONCLUSION

This paper combines a set of methods to overcome shortcomings of current approaches to base macroeconomic models on microfoundations. While representative agent approaches are unable to capture dynamics that emerge from structured and local interactions of multiple heterogeneous agents, computational agent-based approaches have the disadvantage that they make tractable model analysis difficult and computationally challenging. We demonstrated that a combination of approximation techniques allows finding a macrodescription of a multiagent system in which heterogeneous agents interact locally on a complex adaptive network as well as via aggregated quantities. In contrast to previous analytic work, where the network structure was either static [36], restricted to starlike clusters [23], or approximated by a mean-field interaction approach and hence neglected [24,25,29,30,35], we explicitly treat the structure of the adaptive complex interaction network with appropriate approximation methods.

We develop a stylized two-sector investment model, in which investment decisions are driven by a social imitation process, to showcase the three approximations: First, a pair approximation of networked interactions takes into account the heterogeneity in interaction patterns. Second, a moment-closure approximation makes it possible to deal with heterogeneous attributes that characterize the agents. Third, the large-system limit abstracts from effects due to the finite population size. It is only possible to take this limit if the model has at least one of the following properties: (i) individual interactions depend only on relative rather than absolute quantities such that the size of households can be decreased while taking the number of households to infinity or (ii) the economic production functions exhibit constant returns to scale such that they scale linearly with the number of households N. The resulting set of ordinary differential equations captures the effect of local interactions at the system level while still allowing for analytical tractability.

A comparison between a computational version of the ABM and the macrodescription reveals that the approximation works well for parameter values distinct from special cases even if only accounting for first moments. Taking more moments into account would increase the accuracy but comes at the cost of higher dimensionality and complexity of the macroscopic dynamical system.

Our model shows that social imitation dynamics add inertia to the investment decisions in the system that cannot be captured by a representative agent approach. The imitation process results in social learning such that agents tend to direct their investments into the more profitable sector over time. Because of this, the shift of investments from the dirty (fossil) to the clean (renewable) sector is driven only by economic factors, namely, increasing exploration and extraction costs for the fossil energy resource. Thus, we conclude that neutral imitation of better-performing peers is not a feasible mechanism to initiate a bottom-up transformation of the economy. Directed imitation, for example, driven by changes in social norms, and supporting policies that make dirty production less profitable are needed to initiate a transformation towards a sustainable economy in the absence of fossil resource shortage.

Finding a system of ordinary differential equations to approximate ABMs is useful because it makes the analysis of the dynamical properties of the model much easier. One promising application here is bifurcation theory, as illustrated in Sec. IV. Furthermore, it opens the possibility of mathematically proving model properties such as the dependency between different parameters and variables in the model.

In the context of climate economics and policy, the proposed techniques are especially important because they allow investigation of the interplay of learning agents adapting to new policies and effects of shifts in values and preferences. The resulting changes in individual behavior and their impact on macroeconomic dynamics can be studied in a comprehensive modeling framework. Large shifts in investments that are required to reach the goals of the Paris agreement are likely to

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

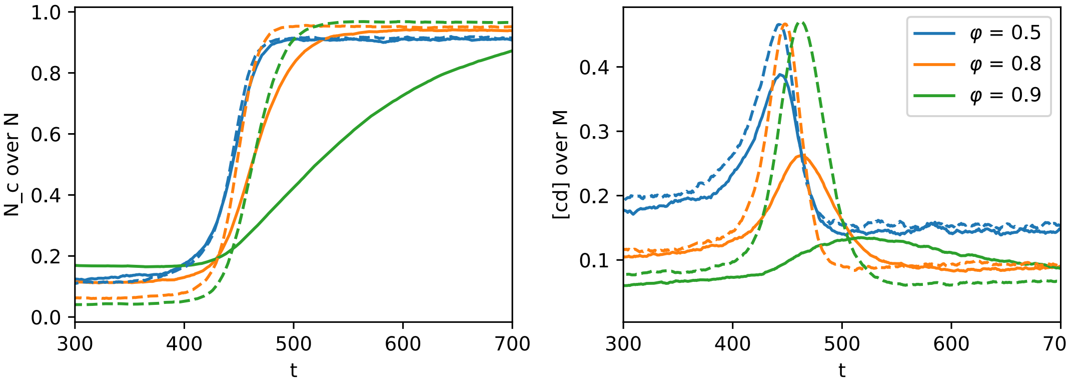


FIG. 6. Comparison of a microscopic model with adaptive network dynamics with a microscopic model with a fully connected network for varying the rewiring rate φ. All other parameters are listed in Table I. Solid lines indicate results with network adaptation; dashed lines, results with a fully connected network. Initial network topology is a Erdős-Renyi random graph.

profit from both policies that rely on price signals and policies that target individual norm change, interaction, and behavior not unlike those researched in, e.g., the public-health context [86,103,104]. The presented techniques can help us to better understand how such behavioral interventions would impact the macrolevel dynamics of the economic system.

In this regard, there are several promising avenues to develop the model and approximation techniques further: For

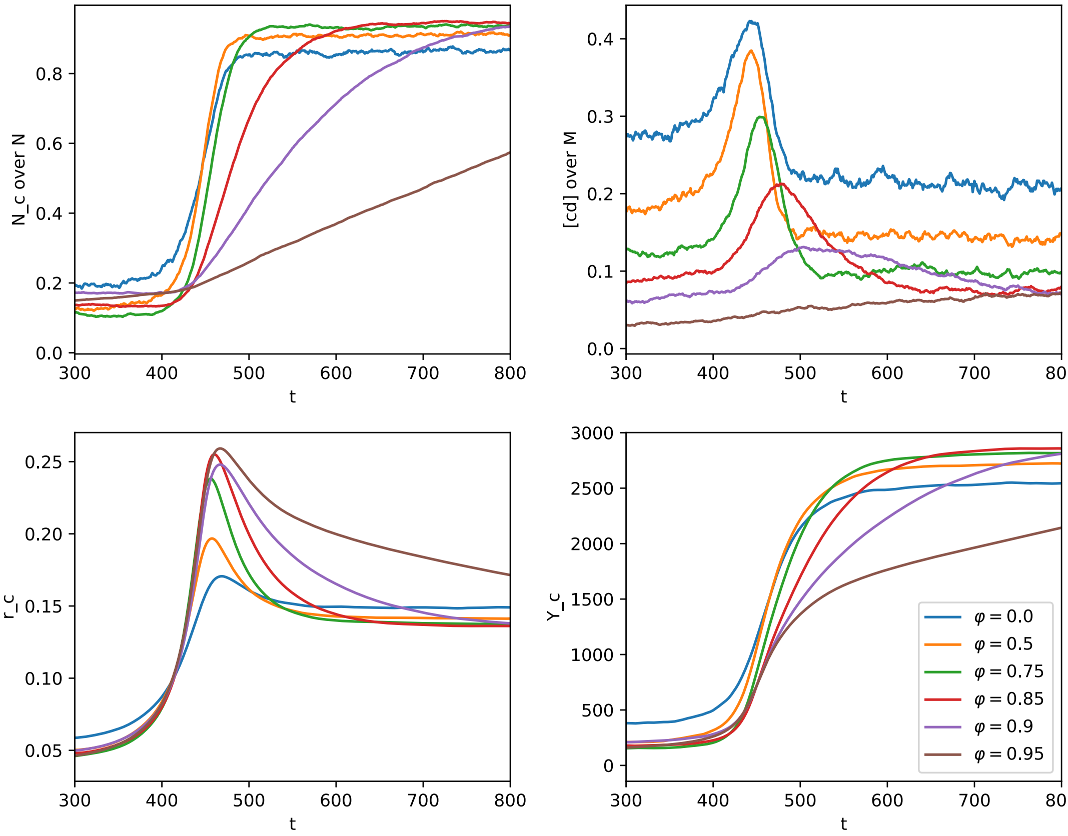


FIG. 7. Model trajectories with varying φ values. All other parameters are listed in Table I. Results are ensemble averages of 200 runs. Initial network topology is a Erdős-Renyi random graph.

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

example, instead of binary opinions, the social interaction model can use continuous variables to represent gradual opinions, drawing on a variety of models of social influence (see Ref. [83], pp. 988 ff.). An approximation of the agent ensemble would then need a Fokker-Planck-type description rather than a master equation.

Our model could be extended to explicitly include policy instruments such as a carbon tax and explore its impact on the investment decisions of the heterogeneous agent population. Another promising modification could include consumption decisions in our two-sector model. Consumption decisions are strongly influenced by social norms and interactions [105]. Their inclusion could inform the discussion about green consumption as a potential mechanism for a bottom-up transformation towards a more sustainable economy.

Finally, the techniques proposed in this paper could be used to approximate other systems that interact both locally in a network and in an aggregate way at the system level, for example, social-ecological systems or neural networks.

ACKNOWLEDGMENTS

The authors declare that they do not have any conflict of interest. J.J.K. acknowledges funding by the Foundation of German Industries. F.M.-H. acknowledges funding by the DFG (IRTG 1740/TRP 2011/50151-0). J.H. acknowledges funding by the Leibniz Assiciation (project DominoES). Numerical analyses were performed on the high-performance computer system of the Potsdam Institute for Climate Impact Research, supported by the European Regional Development Fund, BMBF, and the Land Brandenburg.

APPENDIX A: COMPARING ADAPTIVE WITH FULLY CONNECTED NETWORKS

We compare the dynamics of the micro model with adaptive network rewiring with the dynamics of the micro model with a fully connected acquaintance network. The model with a fully connected acquaintance network is equivalent to a well-mixed model with pairwise interactions between all agents. The results in Fig. 6 show that the well-mixed model approximates the adaptive network model for $\varphi = 0.5$ quite well. However, for increasing φ, the fragmentation increases in the adaptive network model, indicated by the lower fraction of links between agents with different savings decisions (clean and dirty), $[cd]/M$. This cannot be captured by the fully connected network model. As an economically observable result, this leads to significantly slower tipping in the adaptive network model.

APPENDIX B: EFFECTS OF THE REWIRING RATE φ ON MODEL DYNAMICS

We analyze the effect of changes in the network rewiring rate φ on the model dynamics. The results in Fig. 7 indicate that for an increasing rewiring rate φ the model undergoes a transition from a connected network state with a considerable number of connections between agents investing in different sectors to a fragmented network state in which such connections are effectively nonexistent. This transition is especially apparent in the fraction of $[cd]$ links in the network given in Fig. 7(b). This fragmentation transition is well known for adaptive voter-type models [39,41,89,90].

APPENDIX C: ODEs RESULTING FROM APPROXIMATION

The following are the full ordinary differential equations resulting from (50), (51), (28f), and (28i):

$$\dot{x} = -\frac{\epsilon x}{\tau} - \frac{p_{cd} z(\epsilon - 1)(\phi - 1)(x + 1)}{\tau(y + 1)} + \frac{p_{dc} z(\epsilon - 1)(\phi - 1)(x - 1)}{\tau(y - 1)}, \tag{C1}$$

$$\begin{aligned}\dot{y} = &-\frac{m(p_{cd} z(\epsilon - 1)(\phi - 1) - p_{dc} z(\epsilon - 1)(\phi - 1) + 0.5\epsilon(y - 1) + 0.5\epsilon(y + 1))}{\tau}\\ &+\frac{(x - 1)(0.25\epsilon z(x - 1) - 0.25\epsilon(x + 1)(y + z - 1) + 0.5\phi z(\epsilon - 1))}{\tau(y - 1)}\\ &+\frac{(x + 1)(0.25\epsilon z(x + 1) + 0.25\epsilon(x - 1)(y - z + 1) - 0.5\phi z(\epsilon - 1))}{\tau(y + 1)},\end{aligned} \tag{C2}$$

$$\begin{aligned}\dot{z} = &-\frac{\epsilon m(2z - 1)}{\tau} - \frac{0.5 p_{cd} z(\epsilon - 1)(\phi - 1)((x + 1)(y + 1) - 2(y - 2z + 1)(my + m - 0.5x - 0.5))}{\tau(y + 1)^2}\\ &-\frac{0.5 p_{dc} z(\epsilon - 1)(\phi - 1)((x - 1)(y - 1) - 2(y + 2z - 1)(my - m - 0.5x + 0.5))}{\tau(y - 1)^2}\\ &+\frac{(x - 1)(0.25\epsilon z(x - 1) - 0.25\epsilon(x + 1)(y + z - 1) + 0.5\phi z(\epsilon - 1))}{\tau(y - 1)}\\ &-\frac{(x + 1)(0.25\epsilon z(x + 1) + 0.25\epsilon(x - 1)(y - z + 1) - 0.5\phi z(\epsilon - 1))}{\tau(y + 1)},\end{aligned} \tag{C3}$$

$$\dot{K}_c^{(c)} = K_c^{(c)}(-\delta + r_c s) + K_d^{(c)} r_d s + Lsw - \frac{0.5K_c^{(c)}(x+1)(p_{cd}z(\epsilon-1)(\phi-1)+0.5\epsilon(y+1))}{\tau(y+1)} + \frac{0.5K_c^{(d)}(x-1)(p_{dc}z(\epsilon-1)(\phi-1)-0.5\epsilon(y-1))}{\tau(y-1)}, \tag{C4}$$

$$\dot{K}_d^{(d)} = K_d^{(d)}(-\delta + r_d s) + K_c^{(d)} r_c s + Lsw + \frac{0.5K_d^{(c)}(x+1)(p_{cd}z(\epsilon-1)(\phi-1)+0.5\epsilon(y+1))}{\tau(y+1)} - \frac{0.5K_d^{(d)}(x-1)(p_{dc}z(\epsilon-1)(\phi-1)-0.5\epsilon(y-1))}{\tau(y-1)}, \tag{C5}$$

$$\dot{K}_d^{(c)} = -K_d^{(c)}\delta - \frac{0.5K_d^{(c)}(x+1)(p_{cd}z(\epsilon-1)(\phi-1)+0.5\epsilon(y+1))}{\tau(y+1)} + \frac{0.5K_d^{(d)}(x-1)(p_{dc}z(\epsilon-1)(\phi-1)-0.5\epsilon(y-1))}{\tau(y-1)}, \tag{C6}$$

$$\dot{K}_c^{(d)} = -K_c^{(d)}\delta + \frac{0.5K_c^{(c)}(x+1)(p_{cd}z(\epsilon-1)(\phi-1)+0.5\epsilon(y+1))}{\tau(y+1)} - \frac{0.5K_c^{(d)}(x-1)(p_{dc}z(\epsilon-1)(\phi-1)-0.5\epsilon(y-1))}{\tau(y-1)}, \tag{C7}$$

$$\dot{G} = -\frac{L^\pi b_d}{e_R}\left(\frac{\left(b_d\left(K_d^{(c)}+K_d^{(d)}\right)^{\kappa_d}\right)^{\frac{1}{1-\pi}}\left(1-\frac{G_0^2 b_R}{G^2 e_R}\right)^{\frac{1}{1-\pi}}}{\left(b_d\left(K_d^{(c)}+K_d^{(d)}\right)^{\kappa_d}\right)^{\frac{1}{1-\pi}}\left(1-\frac{G_0^2 b_R}{G^2 e_R}\right)^{\frac{1}{1-\pi}} + \left(C^\xi b_c\left(K_c^{(c)}+K_c^{(d)}\right)^{\kappa_c}\right)^{\frac{1}{1-\pi}}}\right)^\pi \left(K_d^{(c)}+K_d^{(d)}\right)^{\kappa_d}, \tag{C8}$$

$$\dot{C} = -C\delta + C^\xi b_c\left(\frac{L\left(C^\xi b_c\left(K_c^{(c)}+K_c^{(d)}\right)^{\kappa_c}\right)^{\frac{1}{1-\pi}}}{\left(b_d\left(K_d^{(c)}+K_d^{(d)}\right)^{\kappa_d}\right)^{\frac{1}{1-\pi}}\left(1-\frac{G_0^2 b_R}{G^2 e_R}\right)^{\frac{1}{1-\pi}} + \left(C^\xi b_c\left(K_c^{(c)}+K_c^{(d)}\right)^{\kappa_c}\right)^{\frac{1}{1-\pi}}}\right)^\pi \times\left(K_c^{(c)}\left(\frac{x}{2}+\frac{1}{2}\right)+K_c^{(d)}\left(\frac{1}{2}-\frac{x}{2}\right)\right)^{\kappa_c}, \tag{C9}$$

where p_{cd} and p_{dc} are given by Eq. (45) and (46) and r_c, r_d, and w are given by

$$r_c = \frac{L^\pi \kappa_c\left(C^\xi b_c\left(K_c^{(c)}+K_c^{(d)}\right)^{\kappa_c}\right)^{\frac{1}{1-\pi}}\left(C^{\frac{1\xi}{1-\pi}} b_c^{\frac{1}{1-\pi}}\left(K_c^{(c)}+K_c^{(d)}\right)^{\frac{1\kappa_c}{1-\pi}} + \left(b_d\left(K_d^{(c)}+K_d^{(d)}\right)^{\kappa_d}\left(1-\frac{G_0^2 b_R}{G^2 e_R}\right)\right)^{\frac{1}{1-\pi}}\right)^{-\pi}}{K_c^{(c)}+K_c^{(d)}}, \tag{C10}$$

$$r_d = \frac{L^\pi \kappa_d\left(b_d\left(K_d^{(c)}+K_d^{(d)}\right)^{\kappa_d}\left(1-\frac{G_0^2 b_R}{G^2 e_R}\right)\right)^{\frac{1}{1-\pi}}}{K_d^{(c)}+K_d^{(d)}} \times\left(C^{\frac{1\xi}{1-\pi}} b_c^{\frac{1}{1-\pi}}\left(K_c^{(c)}+K_c^{(d)}\right)^{\frac{1\kappa_c}{1-\pi}} + \left(b_d\left(K_d^{(c)}+K_d^{(d)}\right)^{\kappa_d}\left(1-\frac{G_0^2 b_R}{G^2 e_R}\right)\right)^{\frac{1}{1-\pi}}\right)^{-\pi}, \tag{C11}$$

$$w = L^{\pi-1}\pi\left(C^{\frac{1\xi}{1-\pi}} b_c^{\frac{1}{1-\pi}}\left(K_c^{(c)}+K_c^{(d)}\right)^{\frac{1\kappa_c}{1-\pi}} + \left(b_d\left(K_d^{(c)}+K_d^{(d)}\right)^{\kappa_d}\right)^{\frac{1}{1-\pi}}\left(1-\frac{G_0^2 b_R}{G^2 e_R}\right)^{\frac{1}{1-\pi}}\right)^{1-\pi}. \tag{C12}$$

[1] V. Grimm and S. F. Railsback, *Individual-Based Modeling and Ecology. Princeton Series in Theoretical and Computational Biology* (Princeton University Press, Princeton, NJ, 2005).

[2] E. Bonabeau, Agent-based modeling: Methods and techniques for simulating human systems, Proc. Natl. Acad. Sci. USA **99**, 7280 (2002).

[3] M. W. Macy and R. Willer, From factors to actors: Computational sociology and agent-based modeling, Annu. Rev. Soc. **28**, 143 (2002).

[4] L. Tesfatsion, Agent-based computational economics: A constructive approach to economic theory, in *Handbook of Computational Economics, Vol. 2*, edited by L. Tesfatsion and K. L. Judd (North-Holland, Amsterdam, 2006), pp. 831–880.

[5] L. Hamill and N. Gilbert, *Agent-Based Modelling in Economics* (Wiley, Chichester, UK, 2016).

[6] J. M. Epstein, Agent-based computational models and generative social science, Complexity **4**, 41 (1999).

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

[7] T. Gross and B. Blasius, Adaptive coevolutionary networks: A review, J. R. Soc. Interf. **5**, 259 (2008).

[8] P. Holme and M. E. J. Newman, Nonequilibrium phase transition in the coevolution of networks and opinions, Phys. Rev. E **74**, 056108 (2006).

[9] L. Bargigli and G. Tedeschi, Interaction in agent-based economics: A survey on the network approach, Phys. A: Stat. Mech. Appl. **399**, 1 (2014).

[10] M. Granovetter, The impact of social structure on economic outcomes, J. Econ. Perspect. **19**, 33 (2005).

[11] D. Delli Gatti, E. Gaffeo, M. Gallegati, G. Giulioni, and A. Palestrini, *Emergent Macroeconomics. An Agent-Based Approach to Business Fluctuations*. *New Economic Windows* (Springer, Milan, 2008).

[12] Approaches to represent heterogeneous agents in DSGE models have been used to counter this criticism and add more realism regarding the distribution of agent attributes [see, for example, the review in [106]. Particularly, because the representative agent approach cannot account for interactions within a heterogeneous group, models using this approach do not allow for the representation of emergent phenomena [107]. However, their solution requires complex numerical methods and cannot integrate local interactions between agents.

[13] A. Kirman, Is it rational to have rational expectations? Mind Soc. **13**, 29 (2014).

[14] G. W. Evans and G. Ramey, Adaptive expectations, underparameterization and the Lucas critique, J. Monet. Econ. **53**, 249 (2006).

[15] We use here a weak notion of emergence, which allows explaining macrophenomena on the basis of microinteractions of the system constituents that differ from the explained macrophenomena. This is opposed to strong emergence, which embraces the irreducibility of macrophenomena to lower-level dynamics. For a discussion see [108].

[16] R. Leombruni and M. Richiardi, Why are economists sceptical about agent-based simulations? Phys. A: Stat. Mech. Appl. **355**, 103 (2005).

[17] V. Grimm, U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. K. Heinz, G. Huse, A. Huth, J. U. Jepsen, C. Jørgensen, W. M. Mooij, B. Müller, G. Pe'er, C. Piou, S. F. Railsback, A. M. Robbins, M. M. Robbins, E. Rossmanith, N. Rüger, E. Strand, S. Souissi, R. A. Stillman, R. Vabø, U. Visser, and D. L. DeAngelis, A standard protocol for describing individual-based and agent-based models, Ecol. Model. **198**, 115 (2006).

[18] J.-S. Lee, T. Filatova, A. Ligmann-Zielinska, B. Hassani-Mahmooei, F. Stonedahl, I. Lorscheid, A. Voinov, G. Polhill, Z. Sun, and D. C. Parker, The complexities of agent-Based modeling output analysis, J. Artif. Soc. Soc. Simulat. **18**, 4 (2015).

[19] R. N. Mantegna and H. E. Stanley, *Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, UK, 1999).

[20] C. Castellano, S. Fortunato, and V. Loreto, Statistical physics of social dynamics, Rev. Mod. Phys. **81**, 591 (2009).

[21] A. D. Martino and M. Marsili, Statistical mechanics of socio-economic systems with heterogeneous agents, J. Phys. A: Math. Gen. **39**, R465 (2006).

[22] D. Acemoglu, A. Ozdaglar, and A. Tahbaz-Salehi, Networks, shocks, and systemic risk, No. 20931, NBER Working Papers, National Bureau of Economic Research, Inc, https://EconPapers.repec.org/RePEc:nbr:nberwo:20931.

[23] C. Di Guilmi, M. Gallegati, S. Landini, and J. E. Stiglitz, Towards an analytical solution for agent based models: An application to a credit network economy, in *Complexity and Institutions: Market Norms and Corporations*, edited by A. Masahiko, B. Kenneth, D. Simon, and G. Herbert (Palgrave Macmillan, New York, 2012), pp. 63–80.

[24] M. Aoki, *New Approaches to Macroeconomic Modeling* (Cambridge University Press, Cambridge, UK, 1996).

[25] M. Aoki and H. Yoshikawa, *Reconstructing Macroeconomics: A Perspective from Statistical Physics and Combinatorial Stochastic Processes* (Cambridge University Press, Cambridge, UK, 2006).

[26] D. Delli Gatti, M. Gallegati, and A. Kirman (Eds.), *Interaction and Market Structure*. *Lecture Notes in Economics and Mathematical Systems* (Springer, Berlin, 2000).

[27] S. Gualdi, M. Tarzia, F. Zamponi, and J. P. Bouchaud, Tipping points in macroeconomic agent-based models, J. Econ. Dyn. Control **50**, 29 (2015).

[28] S. Gualdi, M. Tarzia, F. Zamponi, and J.-P. Bouchaud, Monetary policy and dark corners in a stylized agent-based model, J. Econ. Interact. Coord. **12**, 507 (2017).

[29] C. Di Guilmi, M. Gallegati, and S. Landini, Economic dynamics with financial fragility and mean-field interaction: A model, Phys. A: Stat. Mech. Appl. **387**, 3852 (2008).

[30] C. Chiarella and C. Di Guilmi, The financial instability hypothesis: A stochastic microfoundation framework, J. Econ. Dyn. Control **35**, 1151 (2011).

[31] S. Landini and M. Gallegati, Heterogeneity, interaction and emergence: Effects of composition, Inte. J. Comput. Econ. Econometr. **4**, 339 (2014).

[32] J.-P. Bouchaud, Crises and collective socio-economic phenomena: Simple models and challenges, J. Stat. Phys. **151**, 567 (2013).

[33] D. Fiaschi and M. Marsili, Economic interactions and the distribution of wealth, in *Econophysics and Economics of Games,Social Choices and Quantitative Techniques*, edited by B. Basu, S. Chakravarty, B. Chakrabarti, and K. Gangopadhyay (Springer, Milan, 2010), pp. 61–70.

[34] N. E. Friedkin and E. C. Johnsen, *Social Influence Network Theory* (Cambridge University Press, New York, 2011).

[35] S. Alfarano, T. Lux, and F. Wagner, Time variation of higher moments in a financial market with heterogeneous agents: An analytical approach, J. Econ. Dyn. Control **32**, 101 (2008).

[36] T. Lux, A model of the topology of the bank-firm credit network and its role as channel of contagion, J. Econ. Dyn. Control **66**, 36 (2016).

[37] T. Gross and H. Sayama, Adaptive networks, in *Adaptive Networks* (Springer, Berlin, 2009), pp. 1–8.

[38] A.-L. Do and T. Gross, Contact processes and moment closure on adaptive networks, in *Adaptive Networks: Theory, Models and Applications*, edited by T. Gross and H. Sayama (Springer and NECSI, Cambridge, MA, 2009), pp. 191–208.

[39] G. Demirel, F. Vazquez, G. A. Böhme, and T. Gross, Moment-closure approximations for discrete adaptive networks, Physica D: Nonlin. Phenom.a **267**, 68 (2014).

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

[40] M. Wiedermann, J. F. Donges, J. Heitzig, W. Lucht, and J. Kurths, Macroscopic description of complex adaptive networks co-evolving with dynamic node states, Phys. Rev. E **91**, 052801 (2015).

[41] B. Min and M. S. Miguel, Fragmentation transitions in a coevolving nonlinear voter model, Sci. Rep. **7**, 1 (2017).

[42] C. Kuehn, Moment closure—A brief review, in *Control of Self-Organizing Nonlinear Systems. Understanding Complex Systems*, edited by E. Schöll, S. Klapp, and P. Hövel (Springer, Cham, 2016), pp. 253–271.

[43] IPCC, *Climate Change 2014: Mitigation of Climate Change: Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on* (Cambridge University Press, Cambridge, UK, 2014).

[44] A. Ansar, B. Caldecot, and J. Tibury, *Stranded assets and the fossil fuel divestment campaign: What does divestment mean for the valuation of fossil fuel assets?* (University of Oxford, Oxford, UK, 2013).

[45] L. Mattauch and C. Hepburn, Climate policy when preferences are endogenous and sometimes they are, Midwest Stud. Philos. **40**, 76 (2016).

[46] L. Mattauch, C. Hepburn, and N. Stern, Pigou pushes preferences: Decarbonization and endogenous values (2018).

[47] E. Gsottbauer and J. C. J. M. V. D. Bergh, Environmental policy theory given bounded rationality and other-regarding rreferences, Environ. Resource Econ. **49**, 263 (2011).

[48] H. Hong and M. Kacperczyk, The price of sin: The effects of social norms on markets, J. Financ. Econ. **93**, 15 (2009).

[49] G. Williams, Some determinants of the socially responsible investment decision: A cross-country study, J. Behav. Finance **8**, 43 (2007).

[50] V. Griskevicius, R. B. Cialdini, and N. J. Goldstein, Social norms: An underestimated and underemployed lever for managing climate change, Int. J. Sustain. Commun. **3**, 5 (2008).

[51] T. Masson and I. Fritsche, Adherence to climate change-related ingroup norms: Do dimensions of group identification matter? Eur. J. Soc. Psychol. **44**, 455 (2014).

[52] P. C. Stern, Contributions of psychology to limiting climate change, Am. Psychol. **66**, 303 (2011).

[53] A. Rabinovich, T. A. Morton, and C. C. Duke, Collective self and individual choice: The role of social comparisons in promoting public engagement with climate change, in *Engaging the Public with Climate Change: Behaviour Change and Communication*, edited by L. Whitmarsh, S. O'Neill, and I. Lorenzoni (Earthscan, Oxon, UK, 2011), pp. 66–83.

[54] B. K. Nyborg, J. M. Anderies, A. Dannenberg, T. Lindahl, C. Schill, M. Schlüter, W. N. Adger, K. J. Arrow, S. Barrett, S. Carpenter, F. Stuart, C. Iii, A.-s. Crépin, G. Daily, P. Ehrlich, C. Folke, W. Jager, N. Kautsky, S. A. Levin, O. J. Madsen, S. Polasky, M. Scheffer, E. U. Weber, J. Wilen, A. Xepapadeas, and A. D. Zeeuw, Social norms as solutions, Science **354**, 42 (2016).

[55] A. Bandura, Social learning theory, in *Social learning theory*, edited by A. Bandura and R. H. Walters, Vol. 1 (Prentice-Hall, Englewood Cliffs, NJ, 1977), pp. 1–46.

[56] N. E. Friedkin, Norm formation in social influence networks, Soc. Networks **23**, 167 (2001).

[57] D. Centola, J. C. Gonza, and M. S. Miguel, Homophily, Cultural drift, and the co-evolution of cultural groups, J. Conflict Resolut. **51**, 905 (2007).

[58] D. Kimura and Y. Hayakawa, Coevolutionary networks with homophily and heterophily, Phys. Rev. E **78**, 016103 (2008).

[59] International Monetary Fund, World Economic and Financial Surveys, Tech. Rep. (IMF, Washington, DC, 2011), https://www.elibrary.imf.org/view/IMF081/11381-9781616350598/11381-9781616350598/11381-9781616350598.xml.

[60] R. Hössinger, C. Link, A. Sonntag, J. Stark, R. Hösslinger, C. Link, A. Sonntag, and J. Stark, Estimating the price elasticity of fuel demand with stated preferences derived from a situational approach, Transport. Res. Part A: Policy Pract. **103**, 154 (2017).

[61] X. Labandeira, J. M. Labeaga, and X. López-otero, A meta-analysis on the price elasticity of energy demand, Energy Policy **102**, 549 (2017).

[62] H. E. Daly, Georgescu-Roegen versus Solow/Stiglitz, Ecol. Econ. **22**, 261 (1997).

[63] N. Georgescu-Roegen, Energy and economic myths, South. Econ. J. **41**, 347 (1975).

[64] N. Georgescu-Roegen, Comments on the papers by Daly and Stiglitz, in *Scarcity and Growth Reconsidered*, edited by V. K. Smith (Resources for the Future, New York, 1979), pp. 95–105.

[65] R. U. Ayres, H. Turton, and T. Casten, Energy efficiency, sustainability and economic growth, Energy **32**, 634 (2007).

[66] R. U. Ayres, J. C. J. M. van den Bergh, D. Lindenberger, and B. S. Warr, The underestimated contribution of energy to economic growth, Struct. Change Econ. Dynam. **27**, 79 (2013).

[67] P. Mulder and H. L. F. D. Groot, Structural change and convergence of energy intensity across OECD countries, 1970–2005, Energy Econ. **34**, 1910 (2012).

[68] L. Argote and D. N. Epple, Learning curves in manufacturing, Science **247**, 920 (1990).

[69] T. P. Wright, Factors affecting the cost of airplanes, J. Aeronaut. Sci. **3**, 122 (1936).

[70] B. Nagy, J. D. Farmer, Q. M. Bui, and J. E. Trancik, Statistical basis for predicting technological progress, PLoS ONE **8**, 1 (2013).

[71] S. Kahouli-Brahmi, Technological learning in energy-environment-economy modelling: A survey, Energy Policy **36**, 138 (2008).

[72] P. Dasgupta and G. Heal, The optimal depletion of exhaustible resources, Rev. Econ. Studi. **41**, 3 (1974).

[73] R. Perman, Y. Ma, J. McGilvray, and M. Common, *Natural Resource and Environmental Economics*, 3rd ed. (Pearson Education, London, 2003).

[74] H. A. Simon, Theories of bounded rationality, in *Decision and Organization*, edited by C. B. McGuire and R. Radner (North-Holland, Amsterdam, 1972), pp. 161–176.

[75] H. A. Simon, *Models of Bounded Rationality: Empirically Grounded Economic Reason*, Vol. 3 (MIT Press, Cambridge, MA, 1982).

[76] G. Gigerenzer and R. Selten, *Bounded Rationality: The Adaptive Toolbox* (MIT Press, Cambridge, MA, 2002).

[77] A. Traulsen, D. Semmann, R. D. Sommerfeld, H.-J. Krambeck, and M. Milinski, Human strategy updating in evolutionary games., Proc. Natl. Acad. Sci. USA **107**, 2962 (2010).

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

[78] D. Barkoczi and M. Galesic, Social learning strategies modify the effect of network structure on group performance, Nat. Commun. **7**, 13109 (2016).

[79] D. R. Fisher, J. Waggle, and P. Leifeld, Where does political polarization come from? Locating polarization within the U.S. climate change debate, Am. Behav. Sci. **57**, 70 (2013).

[80] J. Farrell, Corporate funding and ideological polarization about climate change, Proc. Natl. Acad. Sci. USA **113**, 92 (2016).

[81] R. E. Dunlap, A. M. McCright, and J. H. Yarosh, The political divide on climate change: Partisan polarization widens in the U.S., Environ.: Sci. Policy Sustain. Dev. **58**, 4 (2016).

[82] A. M. McCright and R. E. Dunlap, The politicization of climate change and polarization in the American public's views of global warming, 2001–2010,Sociol. Q. **52**, 155 (2011).

[83] P. S. Hart and E. C. Nisbet, Boomerang effects in science communication: How motivated reasoning and identity cues amplify opinion polarization about climate mitigation policies, Commun. Res. **39**, 701 (2012).

[84] H. T. Williams, J. R. McMurray, T. Kurz, and F. H. Lambert, Network analysis reveals open forums and echo chambers in social media discussions of climate change, Glob. Environ. Chang. **32**, 126 (2015).

[85] M. McPherson, L. Smith-Lovin, and J. M. Cook, Birds of a feather: Homophily in social networks, Annu. Rev. Soc. **27**, 415 (2001).

[86] D. Centola, An experimental study of homophily in the adoption of health behavior, Science **334**, 1269 (2011).

[87] Y. M. Asano, J. J. Kolb, J. Heitzig, and J. D. Farmer, Emergent inequality and endogenous dynamics in a simple behavioral macroeconomic model, INET Oxford Working Paper No. 2019-11, 2019 (unpublished).

[88] J. J. Kolb, github.com/jakobkolb/pydivest (2018).

[89] T. Gross, C. J. D. D'Lima, and B. Blasius, Epidemic dynamics on an adaptive network, Phys. Rev. Lett. **96**, 208701 (2006).

[90] G. A. Böhme and T. Gross, Analytical calculation of fragmentation transitions in adaptive networks, Phys. Rev. E **83**, 035101(R) (2011).

[91] T. Rogers, W. Clifford-Brown, C. Mills, and T. Galla, Stochastic oscillations of adaptive networks: Application to epidemic modelling, J. Stat. Mech.: Theory Exp. **2012**, P08018 (2012).

[92] T. Rogers and T. Gross, Consensus time and conformity in the adaptive voter model, Phys. Rev. E **88**, 030102(R) (2013).

[93] J. Nitzbon, J. Heitzig, and U. Parlitz, Sustainability, collapse and oscillations of global climate, population and economy in a simple World-Earth model, Environ. Res. Lett. **12**, 074020 (2017).

[94] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, 2nd ed. (North-Holland Personal Library, Amsterdam, 1992).

[95] F. Müller-Hansen, M. Schlüter, M. Mäs, J. F. Donges, J. J. Kolb, K. Thonicke, and J. Heitzig, Towards representing human behavior and decision making in Earth system models—An overview of techniques and approaches, Earth Syst. Dynam. **8**, 977 (2017).

[96] S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering* (Perseus Books, Reading, MA, 1994).

[97] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Vol. 112, 2nd ed. (Springer Science & Business Media, New York,2013).

[98] E. L. Allgower and K. Georg, *Introduction to Numerical Continuation Methods* (SIAM, Philadelphia, PA, 2003).

[99] L. M. G. J. Clewley, W. E. Sherwood, M. D. LaMar, and J. M. Guckenheimer, PyDSTool, a software environment for dynamical systems modeling (2007); https://pydstool.github.io/PyDSTool/FrontPage.html.

[100] R. Clewley, Hybrid models and biological model reduction with PyDSTool, PLOS Comput. Biol. **8**, e1002628 (2012).

[101] PyDSTool is building on the AUTO-07p continuation library [109].

[102] J. Heitzig, T. Kittel, J. F. Donges, and N. Molkenthin, Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dynam. **7**, 21 (2016).

[103] J. Zhang, D. Brackbill, S. Yang, J. Becker, N. Herbert, and D. Centola, Support or competition? How online social networks increase physical activity: A randomized controlled trial, Prev. Med. Rep. **4**, 453 (2016).

[104] J. Zhang, D. Brackbill, S. Yang, and D. Centola, Efficacy and causal mechanism of an online social media intervention to increase physical activity: Results of a randomized controlled trial, Prev. Med. Rep. **2**, 651 (2015).

[105] K. Peattie, Green consumption: Behavior and norms, Annu. Rev. Environ. Res. **35**, 195 (2010).

[106] J. Heathcote, K. Storesletten, and G. L. Violante, Quantitative macroeconomics with heterogeneous households, Annu. Rev. Econ. **1**, 319 (2009).

[107] A. P. Kirman, Whom or what does the representative individual represent?, J. Econ. Perspect. **6**, 117 (1992).

[108] M. A. Bedau, Weak emergence, Noûs **31**, 375 (1997).

[109] E. J. Doedel, T. F. Fairgrieve, B. Sandstede, A. R. Champneys, Y. A. Kuznetsov, and X. Wang, AUTO-07P: continuation and bifurcation software for ordinary differential equations, 2007; http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.423.2590.

Reprinted paper with permission from J. J. Kolb et al., Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model, Phys. Rev. E, vol. 102, no. 4, 2020, doi: 10.1103/PhysRevE.102.042311. Copyright 2020 by the American Physical Society.

PHYSICAL REVIEW E **99**, 043305 (2019)

Editors' Suggestion

Deterministic limit of temporal difference reinforcement learning for stochastic games

Wolfram Barfuss,[1,2,*] Jonathan F. Donges,[1,3] and Jürgen Kurths[1,2,4]

[1] *Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany*
[2] *Department of Physics, Humboldt University Berlin, 12489 Berlin, Germany*
[3] *Stockholm Resilience Centre, Stockholm University, 104 05 Stockholm, Sweden*
[4] *Saratov State University, 410012 Saratov, Russia*

(Received 20 September 2018; published 10 April 2019)

Reinforcement learning in multiagent systems has been studied in the fields of economic game theory, artificial intelligence, and statistical physics by developing an analytical understanding of the learning dynamics (often in relation to the replicator dynamics of evolutionary game theory). However, the majority of these analytical studies focuses on repeated normal form games, which only have a single environmental state. Environmental dynamics, i.e., changes in the state of an environment affecting the agents' payoffs has received less attention, lacking a universal method to obtain deterministic equations from established multistate reinforcement learning algorithms. In this work we present a methodological extension, separating the interaction from the adaptation timescale, to derive the deterministic limit of a general class of reinforcement learning algorithms, called temporal difference learning. This form of learning is equipped to function in more realistic multistate environments by using the estimated value of future environmental states to adapt the agent's behavior. We demonstrate the potential of our method with the three well-established learning algorithms Q learning, SARSA learning, and actor-critic learning. Illustrations of their dynamics on two multiagent, multistate environments reveal a wide range of different dynamical regimes, such as convergence to fixed points, limit cycles, and even deterministic chaos.

DOI: 10.1103/PhysRevE.99.043305

I. INTRODUCTION

Individual learning through reinforcements is a central approach in the fields of artificial intelligence [1–3], neuroscience [4,5], learning in games [6], and behavioral game theory [7–10], thereby offering a general purpose principle to either solve complex problems or explain behavior. Also in the fields of complexity economics [11,12] and social science [13], reinforcement learning has been used as a model for human behavior to study social dilemmas.

However, there is a need for improved understanding and better qualitative insight into the characteristic dynamics that different learning algorithms produce. Therefore, reinforcement learning has also been studied from a dynamical systems perspective. In their seminal work, Börgers and Sarin showed that one of the most basic reinforcement learning update schemes, Cross learning [14], converges to the replicator dynamics of evolutionary games theory in the continuous time limit [15]. This has led to at least two, presumably nonoverlapping, research communities, one from statistical physics [16–26] and one from computer science machine learning [27–35]. Thus, Sato and Crutchfield [18] and Tuyls *et al.* [27] independently deduced identical learning equations in 2003.

The statistical physics articles usually consider the deterministic limit of the stochastic learning equations, assuming infinitely many interactions between the agents before an adaptation of behavior occurs. This limit can either be performed in continuous time with differential equations [17–19] or discrete time with difference equations [20–22]. The differences between both variants can be significant [21,23]. Deterministic chaos was found to emerge when learning simple [17] as well as complicated games [25]. Relaxing the assumption of infinitely many interactions between behavior updates revealed that noise can change the attractor of the learning dynamics significantly, e.g., by noise-induced oscillations [20,21].

However, these statistical physics studies so far considered only repeated normal form games. These are games where the payoff depends solely on the set of current actions, typically encoded in the entries of a payoff matrix (for the typical case of two players). Receiving payoff and choosing another set of joint actions is performed repeatedly. This setup lacks the possibility to study dynamically changing environments and their interplay with multiple agents. In those systems, rewards depend not only on the joint action of agents but also on the states of the environment. Environmental state changes may occur probabilistically and depend also on joint actions and the current state. Such a setting is also known as a Markov game or stochastic game [36,37]. Thus, a repeated normal form game is a special case of a stochastic game with only one environmental state. Notably, Akiyama and Kaneko [38,39] did emphasize the importance of a dynamically changing environment; however, they did not utilize a reinforcement learning update scheme.

The computer science machine-learning community dealing with reinforcement learning as a dynamical system (see Ref. [28] for an overview) particularly emphasizes the link between evolutionary game theory and multiagent reinforcement learning as a well grounded theoretical framework for the

[*] barfuss@pik-potsdam.de

latter [28–31]. This dynamical systems perspective is proposed as a way to gain qualitative insights about the variety of multiagent reinforcement learning algorithms (see Ref. [2] for a review). Consequently, this literature developed a focus on the translation of established reinforcement learning algorithms to a dynamical systems description, as well as the development of new algorithms based on insights of a dynamical systems perspective. While there is more work on stateless games (e.g., Q learning [27] and frequency-adjusted multiagent Q learning [32]), multiagent learning dynamics for multistate environments have been developed as well, such as piecewise replicator dynamics [34], state-coupled replicator dynamics [33], or reverse engineering state-coupled replicator dynamics [35].

Both communities, statistical physics and machine learning, share the interest in better qualitative insights into multiagent learning dynamics. While the statistical physics community focuses more on dynamical properties the same set of learning equations can produce, it leaves a research gap of learning equations capable of handling multiple environmental states. The machine-learning community, on the other hand, aims more toward algorithm development, but so far have put their focus less on a dynamical systems understanding. Taken together, there is the challenge of developing a dynamical systems theory of multiagent learning dynamics in varying environmental states.

With this work, we aim to contribute to such a dynamical systems theory of multiagent learning dynamics. We present a methodological extension for obtaining the deterministic limit of multistate temporal difference reinforcement learning. In essence, it consists of formulating the temporal difference error for batch learning, and sending the batch size to infinity. We showcase our approach with the three prominent learning variants of Q learning, SARSA learning, and actor-critic (AC) learning. Illustrations of their learning dynamics reveal multiple different dynamical regimes, such as fixed points, periodic orbits, and deterministic chaos.

In Sec. II we introduce the necessary background and notation. Section III presents our method to obtain the deterministic limit of temporal difference reinforcement learning and demonstrates it for multistate Q learning, SARSA learning, and actor-critic learning. We illustrate their learning dynamics for two previously utilized two-agent two-action two-state environments in Sec. IV. In Sec. V we conclude with a discussion of our work.

II. PRELIMINARIES

We introduce the components (including notation) of our multiagent environment systems (see Fig. 1), followed by a brief introduction of temporal difference reinforcement learning.

A. Multi-agent Markov environments

A multiagent Markov environment (also called stochastic game or Markov game) consists of $N \in \mathbb{N}$ *agents*. The environment can exist in $Z \in \mathbb{N}$ *states* $\mathcal{S} = \{S_1, \dots, S_Z\}$. In each state each agent has $M \in \mathbb{N}$ available *actions* $\mathcal{A}^i = \{A^i_1, \dots, A^i_M\},\ i = 1, \dots, N$ to choose from. Having an identical number of actions for all states and all agents is notational convenience, no significant restriction. A joint action of all agents is referred to by $\mathbf{a} \in \boldsymbol{\mathcal{A}} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$, the joint action of all agents but agent i is denoted by $\mathbf{a}^{-i} \in \boldsymbol{\mathcal{A}}^{-i} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^{i-1} \times \mathcal{A}^{i+1} \times \cdots \times \mathcal{A}^N$.

Environmental dynamics are given by the probabilities for state changes expressed as a transition tensor $\mathbf{T} \in [0, 1]^{Z\times M\times\dots(\text{N times})\dots\times M\times Z}$. The entry $T_{s\mathbf{a}s'}$ denotes the probability $P(s'|s, \mathbf{a})$ that the environment transitions to state s' given the environment was in state s and the agents have chosen the joint action $\mathbf{a}$. Hence, for all $s, \mathbf{a}$, $\sum_{s'} T_{s\mathbf{a}s'} = 1$ must hold. The assumption that the next state only depends on the current state and joint action makes our system Markovian. We here restrict ourselves to ergodic environments without absorbing states (cf. Ref. [35]).

The rewards receivable by the agents are given by the reward tensor $\mathbf{R} \in \mathbb{R}^{N\times Z\times M\times\dots(\text{N times})\dots\times M\times Z}$. The entry $R^i_{s\mathbf{a}s'}$ denotes the reward agent i receives when the environment transits from state s to state s' under the joint action $\mathbf{a}$. Rewards are also called payoffs from a game-theoretic perspective.

Agents draw their actions from their behavior profile $\mathbf{X} \in [0, 1]^{N\times Z\times M}$. The entry $X^i_{sa} = P(a\,|\,i, s)$ denotes the probability that agent i chooses action a in state s. Thus, for all i and all s, $\sum_a X^i_{sa} = 1$ must hold. We here focus on the case of independent agents, able to fully observe the current state of the environment. With correlated behavior (see, e.g., Ref. [2]) and partially observable environments [40,41], one could extend the multiagent environment systems to be even more general. Note that what we call behavior profile is usually termed policy from a machine-learning perspective or behavioral strategy from a game-theoretic perspective. We chose to introduce our own term because policies and strategies suggest a deliberate choice which we do not want to impose.

B. Averaging out behavior and environment

We define a notational convention that allows a systematic averaging over the current behavior profile $\mathbf{X}$ and the environmental transitions $\mathbf{T}$. It will be used throughout the paper.

Averaging over the whole behavioral profile yields

$$\begin{aligned}\mathbf{x}\langle\circ\rangle &:= \sum_{\mathbf{a}} \mathbf{X}_{s\mathbf{a}} \cdot \circ \\ &:= \sum_{a^1\in\mathcal{A}^1} \cdots \sum_{a^N\in\mathcal{A}^N} X^1_{sa^1} \cdots X^N_{sa^N} \cdot \circ. \qquad (1)\end{aligned}$$

Here $\circ$ serves as a placeholder. If the quantity to be inserted for $\circ$ depends on the summation indices, then those indices will be summed over as well. If the quantity, which is averaged out, is used in tensor form, then it is written in bold. If not, then remaining indices are added after the right angle bracket.

Averaging over the behavioral profile of the other agents, keeping the action of agent i, yields

$$\begin{aligned}\mathbf{x}_{-i}\langle\circ\rangle &:= \sum_{\mathbf{a}^{-i}} \mathbf{X}^{-i}_{s\mathbf{a}^{-i}} \cdot \circ \\ &:= \underbrace{\sum_{a^1\in\mathcal{A}^1} \cdots \sum_{a^N\in\mathcal{A}^N}}_{\text{excl. } i} \underbrace{X^1_{sa^1} \cdots X^N_{sa^N}}_{\text{excl. } i} \cdot \circ\,. \qquad (2)\end{aligned}$$

Reprinted paper with permission from W. Barfuss et al., Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E, vol. 99, no. 4, pp. 1-16, 2019, doi: 10.1103/PhysRevE.99.043305. Copyright 2019 by the American Physical Society

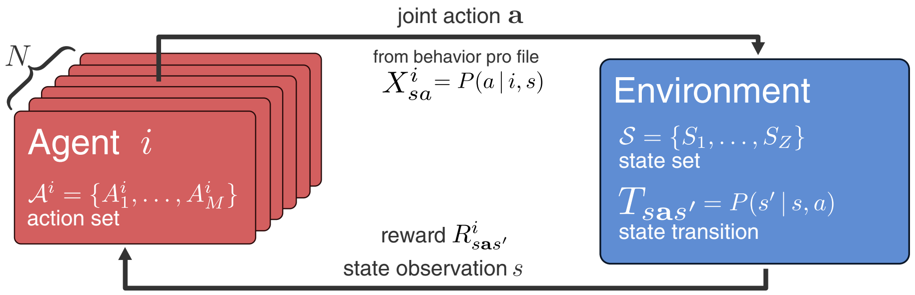

joint action $\mathbf{a}$
N
from behavior pro file
$X^i_{sa} = P(a\,|\,i, s)$
Agent i
$\mathcal{A}^i = \{A^i_1, \dots, A^i_M\}$
action set
Environment
$\mathcal{S} = \{S_1, \dots, S_Z\}$
state set
$T_{s\mathbf{a}s'} = P(s'\,|\,s, a)$
state transition
reward $R^i_{s\mathbf{a}s'}$
state observation s

FIG. 1. Multiagent Markov environment (also known as stochastic or Markov game). N agents choose a joint action $\mathbf{a} = (a^1, \dots, a^N)$ from their action sets $\mathcal{A}^i$, based on the current state of the environment s, according to their behavior profile $X^i_{sa} = P(a|i, s)$. This will change the state of the environment from s to s' with probability $T_{s\mathbf{a}s'}$ and provide each agent with a reward $R^i_{s\mathbf{a}s'}$.

Last, averaging over the subsequent state s' yields

$$ {}_{\mathbf{T}}\langle\circ\rangle := \sum_{s'} T_{s\mathbf{a}s'} \cdot \circ := \sum_{s' \in \mathcal{S}} T_{sa^1 \dots a^N s'} \cdot \circ. \tag{3}$$

Of course, these operations may also be combined as ${}_{\mathbf{TX}}\langle\circ\rangle$ and ${}_{\mathbf{TX}^{-i}}\langle\circ\rangle$ by multiplying both summations.

For example, given a behavior profile $\mathbf{X}$, the resulting effective Markov Chain transition matrix reads ${}_{\mathbf{X}}\langle T\rangle_{ss'}$, which encodes the transition probabilities from state s to s'. From ${}_{\mathbf{X}}\langle T\rangle_{ss'}$ the stationary distribution of environmental states $\boldsymbol{\sigma}(\mathbf{X})$ can be computed. $\boldsymbol{\sigma}(\mathbf{X})$ is the eigenvector corresponding to the eigenvalue 1 of ${}_{\mathbf{X}}\langle T\rangle_{ss'}$. Its entries encode the ratios of the average durations the agents find themselves in the respective environmental states.

The average reward agent i receives from state s under action a, given all other agents follow the behavior profile $\mathbf{X}$ reads ${}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}$. Including agent i's behavior profile gives the average reward it receives from state s: ${}_{\mathbf{TX}}\langle R\rangle^i_s$. Hence, ${}_{\mathbf{TX}}\langle R\rangle^i_s = \sum_a X^i_{sa} \cdot {}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}$ holds.

C. Agent's preferences and values

Typically, agents are assumed to maximize their exponentially discounted sum of future rewards, called return $G^i(t) = (1 - \gamma^i) \sum_{k=0}^{\infty} (\gamma^i)^k r^i(t + k)$, where $\gamma^i \in [0, 1)$ is the discount factor of agent i and $r^i(t + k)$ denotes the reward received by agent i at time step $t + k$. Exponential discounting is most commonly used for its mathematical convenience and because it ensures consistent preferences over time. Other formulations of a return use, e.g., finite-time horizons, average reward settings, as well as other ways of discounting, such as hyperbolic discounting. Those other forms require their own form of reinforcement learning.

Given a behavior profile $\mathbf{X}$, the expected return defines the *state-value function* $V^i_s(\mathbf{X}) := {}_{\mathbf{TX}}\langle G^i(t) \mid s(t) = s\rangle^i_s$, which is independent of time t. The operation ${}_{\mathbf{TX}}\langle \dots \mid s(t) = s\rangle$ denotes the behavioral and environmental average as defined in Eqs. (1) and (3) given that in the current time step t the environment is in state s. Inserting the return yields the *Bellman equation* [42],

$$V^i_s(\mathbf{X}) = {}_{\mathbf{TX}}\Big\langle (1 - \gamma^i) r^i(t) + \gamma^i V^i_{s(t+1)}(\mathbf{X}) \Big| s(t) = s \Big\rangle^i_s. \tag{4}$$

This recursive relationship between state values declares that the value of a state s is the discounted value of the subsequent state $s(t + 1)$ plus $(1 - \gamma^i)$ times the reward received along the way. Evaluating the behavioral and environmental average ${}_{\mathbf{TX}}\langle\ \rangle$ and writing in matrix form we get:

$$\mathbf{V}^i(\mathbf{X}) = (1 - \gamma^i) \cdot {}_{\mathbf{TX}}\langle \mathbf{R}\rangle^i + \gamma^i \cdot {}_{\mathbf{X}}\langle \mathbf{T}\rangle \cdot \mathbf{V}^i(\mathbf{X}). \tag{5}$$

The reward $r^i(t)$ received at time step t is evaluated to reward ${}_{\mathbf{TX}}\langle R\rangle^i_s$ for state s, since the behavioral and environmental average was conditioned on starting in state $s(t) = s$. The average subsequent state value $V^i_{s(t+1)}(\mathbf{X})$ from the current state s can be expressed as a matrix multiplication of the effective Markov transition matrix and the vector of state values: $\sum_{s'} {}_{\mathbf{X}}\langle T\rangle_{ss'} \cdot \mathbf{V}^i_{s'}(\mathbf{X})$.

A solution of the state values $\mathbf{V}^i(\mathbf{X})$ can be obtained using matrix inversion

$$\mathbf{V}^i(\mathbf{X}) = (1 - \gamma^i)(\mathbb{1}_Z - \gamma^i {}_{\mathbf{X}}\langle \mathbf{T}\rangle)^{-1} {}_{\mathbf{TX}}\langle \mathbf{R}\rangle^i. \tag{6}$$

The computational complexity of matrix inversion makes this solution strategy infeasible for large systems. Therefore many iterative solution methods exist [3].

Equivalently, *state-action-value functions* Q^i_{sa} are defined as the expected return, given agent i applied action a in state s and then followed $\mathbf{X}$ accordingly: $Q^i_{sa}(\mathbf{X}) := {}_{\mathbf{TX}}\langle G^i(t) \mid s(t) = s, a(t) = a\rangle^i_{sa}$. Even though this is the behavioral average over the whole behavioral profile, the resulting object carries an action index because the operation is conditioned on the current action to be $a(t) = a$. They can be computed via

$$Q^i_{sa}(\mathbf{X}) = (1 - \gamma^i) {}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^i \sum_{s'} {}_{\mathbf{X}}\langle T\rangle_{ss'} \cdot V^i_{s'}(\mathbf{X}). \tag{7}$$

One can show that $V^i_s(\mathbf{X}) = \sum_a X^i_{sa} Q^i_{sa}(\mathbf{X})$ holds for the inverse relation of state-action and state values.

D. Learning through reinforcement

In contrast to the typical game-theoretic assumption of perfect information, we assume that agents know nothing about the game in advance. They can only gain information about the environment and other agents through interactions. They do not know the true reward tensor $\mathbf{R}$ or the true transition probabilities $T_{s\mathbf{a}s'}$. They experience only reinforcements (i.e., particular rewards $R^i_{s\mathbf{a}s'}$), while observing the current true Markov state of the environment.

Reprinted paper with permission from W. Barfuss et al., Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E, vol. 99, no. 4, pp. 1-16, 2019, doi: 10.1103/PhysRevE.99.043305. Copyright 2019 by the American Physical Society

In essence, reinforcement learning consists of iterative behavior changes toward a behavior profile with maximum state values. However, due to the agents' limited information about the environment, they generally cannot compute a behavior profile's true state and state-action values, $V^i_s(\mathbf{X})$ and $Q^i_{sa}(\mathbf{X})$, as defined in the previous section. Therefore, agents use time-dependent *state-value* and *state-action-value approximations*, $\tilde{V}^i_s(t)$ and $\tilde{Q}^i_{sa}(t)$, during the reinforcement learning process.

1. *Temporal difference learning*

Basically, state-action-value approximations $\tilde{Q}^i_{sa}$ get iteratively updated by a temporal difference error $D^i_{sa}(t)$:

$$\tilde{Q}^i_{sa}(t+1) = \tilde{Q}^i_{sa}(t) + \alpha^i D^i_{sa}(t), \tag{8}$$

with $\alpha^i \in (0, 1)$ being the *learning rate* of agent i. These state-action propensities $\tilde{Q}^i_{sa}$ can be interpreted as estimates of the state-action values Q^i_{sa}.

The temporal difference error expresses a difference in the estimation of state-action values. New experience is used to compute a new estimate of the current state-action value and corrected by the old estimate. The estimate from the new experience uses exactly the recursive relation of value functions from the Bellmann equation [Eq. (4)],

$$\begin{aligned} D^i_{sa}(t) = {} & \delta_{ss(t)}\delta_{aa(t)} \\ & \cdot\Bigg[\underbrace{(1-\gamma^i)R^i_{s(t)a(t)\mathbf{a}^{-i}(t)s(t+1)} + \gamma^i\, \Upsilon^i_{s(t+1)}(t)}_{\text{estimate from new experience}} \\ & - \underbrace{\Upsilon^i_{s(t)}(t)}_{\text{old estimate}}\Bigg]. \end{aligned} \tag{9}$$

Here s and a denote the state-action pair whose temporal difference error is calculated. With $s(t)$, $a(t)$, etc., we refer to the state, action, etc., that occurred at time step t. Thus, the notation $R^i_{s(t)a(t)\mathbf{a}^{-i}(t)s(t+1)}$ refers to the entry of the reward tensor $R^i_{sa\mathbf{a}^{-i}s'}$ when at time step t the environmental state was s [$s(t) = s$], agent i chose action a [$a(t) = a$], the other agents chose the joint action $\mathbf{a}^{-i}$ [$\mathbf{a}^{-i}(t) = \mathbf{a}^{-i}$] and the next environmental state was s' [$s(t+1) = s'$]. The $\Upsilon^i_{s(t+1)}(t)$ indicates the state-value estimate at time step t of the state visited at the next time step $s(t+1)$. $\Upsilon^i_{s(t)}(t)$ denotes the state-value estimate at time step t of the current state $s(t)$. Different choices for these estimations are possible, leading to different learning variants (see below).

The Kronecker deltas $\delta_{ss(t)}$, $\delta_{aa(t)}$ indicate that the temporal difference error for state-action pair (s, a) is only nonzero when (s, a) was actually visited in time step t. This denotes and emphasizes that agents can only learn from experience. In contrast, e.g., experience-weighted-attraction learning [9] assumes that action propensities can be updated with hypothetical rewards an agent would have received if she had played a different action than the current action. These two cases have been referred to as *full vs. partial information* [16]. Thus, the Kronecker deltas in Eq. (9) indicate a partial information update. The agents use only information experienced through interaction.

The state-action-value approximations $\tilde{Q}^i_{sa}$ are translated to a behavior profile according to the Gibbs-Boltzmann distribution [1] (also called softmax),

$$X^i_{sa}(t) = \frac{\exp\left[\beta^i \tilde{Q}^i_{sa}(t)\right]}{\sum_b \exp\left[\beta^i \tilde{Q}^i_{sb}(t)\right]}. \tag{10}$$

The behavior profile $\mathbf{X}$ becomes a dynamic variable as well. The parameter β^i controls the *intensity of choice* or the *exploitation level* of agent i controlling the *exploration-exploitation trade-off*. In analogy to statistical physics, β^i is the inverse temperature. For high β^i, agents tend to exploit their learned knowledge about the environment, leaning toward actions with high estimated state-action value. For low β^i, agents are more likely to deviate from these high-value actions in order to explore the environment further with the chance of finding actions, which eventually lead to even higher values. Other behavior profile translations exist as well (e.g., ϵ-greedy [1]).

2. *Three learning variants*

The specific choices of the value estimates Υ in the temporal difference error result in different reinforcement learning variants.

a. Q learning. For the Q learning algorithm [1,3], $\Upsilon^i_{s(t+1)}(t) = \max_b \tilde{Q}^i_{s(t+1)b}(t)$ and $\Upsilon^i_{s(t)}(t) = \tilde{Q}^i_{s(t)a(t)}(t)$. Thus, the Q learning update takes the maximum of the next state-action-value approximations as an estimate for the next state value, regardless of the actual next action the agent plays. This is reasonable because the maximum is the highest value achievable given the current knowledge. For the state-value estimate of the current state, the Q learner takes the current state-action-value approximation $Q^i_{s(t)a(t)}(t)$. This is reasonable because it is exactly the quantity that gets updated by Eq. (8).

b. SARSA learning. For SARSA learning [1,3], $\Upsilon^i_{s(t+1)}(t)= \tilde{Q}^i_{s(t+1)a(t+1)}(t)$ and $\Upsilon^i_{s(t)}(t) = \tilde{Q}^i_{s(t)a(t)}(t)$, where $a(t+1)$ denotes the action taken by agent i at the next time step. Thus, the SARSA algorithm uses the five ingredients of an update sequence of state, action, reward, next state, and next action to perform one update. In practice, the SARSA sequence has to be shifted one time step backward to know what the actual "next" action of the agent was.

c. Actor-critic learning. For AC learning [1,3], $\Upsilon^i_{s(t+1)}(t)= \tilde{V}^i_{s(t+1)}(t)$ and $\Upsilon^i_{s(t)}(t) = \tilde{V}^i_{s(t)}(t)$. Compared to Q and SARSA learners, it has an additional data structure of state-value approximations which get separately updated according to $\tilde{V}^i_s(t+1) = \tilde{V}^i_s(t) + \alpha^i \cdot D^i_{sa}(t)$. The state-action-value approximations $\tilde{Q}^i_{sa}$ serve as the actor which gets criticized by the state-value approximations $\tilde{V}^i_s$.

Table I summarizes the values estimates Υ for these three learning variants. Q and SARSA learning are structurally more similar compared to the actor-critic learner, which uses an additional data structure of state-value approximations $\tilde{V}^i_s$.

III. DETERMINISTIC LIMIT

So far we gave a brief introduction to temporal difference reinforcement learning. A more comprehensive presentation can be found in Ref. [1]. In this section we will present an extension to the methodology of interaction-adaptation timescales separation to the general class of temporal

Reprinted paper with permission from W. Barfuss et al., Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E, vol. 99, no. 4, pp. 1-16, 2019, doi: 10.1103/PhysRevE.99.043305. Copyright 2019 by the American Physical Society

TABLE I. Overview of the three reinforcement learning variants. Shown in the columns are the value estimates for the next state $\Upsilon^i_{s(t+1)}(t)$ and the current state $\Upsilon^i_{s(t)}(t)$ for both ends of the batch size spectrum: $K = 1$ and $K = \infty$.

	(a) $K = 1$		(b) $K = \infty$	
	$\Upsilon^i_{s(t+1)}(t)$	$\Upsilon^i_{s(t)}(t)$	$\Upsilon^i_{s(t+1)}(t)$	$\Upsilon^i_{s(t)}(t)$
Q learning	$\max_b \tilde{Q}^i_{s(t+1)b}(t)$	$\tilde{Q}^i_{s(t)a(t)}(t)$	${}^{\max}Q^i_{sa}(\mathbf{X})$	$\frac{1}{\beta^i}\log X^i_{sa}(t)$
SARSA learning	$\tilde{Q}^i_{s(t+1)a(t+1)}(t)$	$\tilde{Q}^i_{s(t)a(t)}(t)$	${}^{\text{next}}\mathcal{V}^i_{sa}(\mathbf{X})$	$\frac{1}{\beta^i}\log X^i_{sa}(t)$
AC learning	$\tilde{V}^i_{s(t+1)}(t)$	$\tilde{V}^i_{s(t)}(t)$	${}^{\text{next}}\mathcal{V}^i_{sa}(\mathbf{X})$	/

difference reinforcement learning. In summary, we (i) give a batch formulation of the temporal difference error, (ii) separate the timescales of interaction and adaptation by sending the batch size to infinity, and (iii) present a resulting deterministic limit conversion rule for discrete time updates. We showcase our method in the three learning variants of Q, SARSA, and actor-critic learning. For the statistical physics community, we present learning equations, capable of handling environmental state transitions. For the machine-learning community, we present the systematic methodology we use to obtain the deterministic learning equations. Note that these deterministic learning equations will not depend on the state-value or state-action-value approximations anymore, being iterated maps of the behavior profile alone.

Following, e.g., Refs. [18,19,22], we first combine Eqs. (8) and (10) and obtain

$$X^i_{sa}(t+1) = \frac{X^i_{sa}(t)\exp\left[\alpha^i\beta^i D^i_{sa}(t)\right]}{\sum_b X^i_{sb}(t)\exp\left[\alpha^i\beta^i D^i_{sb}(t)\right]}. \tag{11}$$

Although it appears that only the product $\alpha^i\beta^i$ matters for a behavior profile update, the temporal difference error D^i_{sa} may depend only on the exploitation level β^i, as we will show below.

Next, we formulate the temporal difference error for batch learning.

A. Batch learning

With batch learning we mean that several time steps of interaction with the environment and the other agents take place before an update of the state-action-value approximations and the behavior profile occurs. It has also been interpreted as a form of history replay [43] which is essential to stabilize the learning process when function approximation (e.g., by deep neural networks) is used [44]. History (i.e., already experienced state, action, next state triples) is used again for an update of the state-action-value approximations.

Imagine that the information from these interactions are stored inside a batch of size $K \in \mathbb{N}$. We introduce the corresponding temporal difference error of batch size K:

$$\begin{aligned} D^i_{sa}(t;K) := \frac{1}{K(s,a)} \sum_{k=0}^{K-1} \big\{ & \delta_{ss(t+k)}\delta_{aa(t+k)} \\ & \times \left[(1-\gamma^i)R^i_{s(t+k)a(t+k)\mathbf{a}^{-i}(t+k)s(t+k+1)} \right. \\ & \left. + \gamma^i\, \Upsilon^i_{s(t+k+1)}(t) - \Upsilon^i_{s(t)}(t)\right]\big\}, \end{aligned} \tag{12}$$

where $K(s,a) = \max[1, \sum_{k=0}^{K-1}\delta_{ss(t+k)}\delta_{aa(t+k)}]$ denotes the number of times the state-action pair (s, a) was visited. If the state-action pair (s, a) was never visited, then $K(s, a) = 1$. The agents interact K times under the same behavior profile and use the sample average to summarize the new experience in order to update the state-action-value approximations:

$$\tilde{Q}^i_{sa}(t+K) = \tilde{Q}^i_{sa}(t) + \alpha^i D^i_{sa}(t;K). \tag{13}$$

The notation $D^i_{sa}(t)$ denotes a batch update of batch size 1: $D^i_{sa}(t) = D^i_{sa}(t;1)$.

B. Separation of timescales

We obtain the deterministic limit of the temporal difference learning dynamics by sending the batch size to infinity, $K \to \infty$. Equivalently, this can be regarded as a separation of timescales. Two processes can be distinguished during an update of the state-action-value approximations $\Delta\tilde{Q}^i_{sa}(t) := \tilde{Q}^i_{sa}(t+1) - \tilde{Q}^i_{sa}(t)$: adaptation and interaction,

$$\Delta\tilde{Q}^i_{sa}(t) = \alpha^i\delta_{ss(t)}\delta_{aa(t)} \cdot \overbrace{\Big[\underbrace{(1-\gamma^i)R^i_{s(t)a(t)\mathbf{a}^{-i}(t)s(t+1)} + \gamma^i\,\Upsilon^i_{s(t+1)}(t)}_{\text{interaction}} - \Upsilon^i_{s(t)}(t)\Big]}^{\text{adaptation}}. \tag{14}$$

By separating the timescales of both processes, we assume that (infinitely) many interactions happen before one step of behavior profile adaptation occurs.

Under this assumption and because of the assumed ergodicity one can replace the sample average, i.e., the sum over sequences of states and actions with the behavior profile average, i.e., the sum over state-action behavior and transition probabilities according to

$$\boxed{\frac{1}{K(s,a)}\sum_{k=0}^{K-1}\delta_{ss(t+k)}\delta_{aa(t+k)} \to \sum_{s'}\sum_{\mathbf{a}^{-i}} \mathbf{X}^{-i}_{\mathbf{sa}^{-i}} T_{sa\mathbf{a}^{-i}s'}.} \tag{15}$$

For example, the immediate reward $R^i_{s(t)a(t)\mathbf{a}^{-\mathbf{i}}(t)s(t+1)}$ in the temporal difference error becomes ${}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}$. The time t gets resealed accordingly, as well.

Taking the limit $K \to \infty$ in this way, we choose to stay in discrete time, leaving the continuous time limit following Refs. [18,19,25] for future work.

Reprinted paper with permission from W. Barfuss et al., Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E, vol. 99, no. 4, pp. 1-16, 2019, doi: 10.1103/PhysRevE.99.043305. Copyright 2019 by the American Physical Society

C. Three learning variants

Next we present the deterministic limit of the temporal difference error of the three learning variants of Q, SARSA, and actor-critic learning. Inserting them into Eq. (11) yields the complete description of the behavior profile update in the deterministic limit. Table I presents an overview of the resulting equations and a comparison to their batch size $K = 1$ versions.

1. Q learning

The temporal difference error of Q learning consists of three terms: (i) $R^i_{s(t)a(t)\mathbf{a}^{-\mathbf{i}}(t)s(t+1)}$, (ii) $\max_b \tilde{Q}^i_{s(t+1)b}(t)$, and (iii) $\tilde{Q}^i_{s(t)a(t)}(t)$. As already stated, $R^i_{s(t)a(t)\mathbf{a}^{-\mathbf{i}}(t)s(t+1)} \rightarrow {}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}$ under $K \rightarrow \infty$. $\max_b \tilde{Q}^i_{s(t+1)b}(t) \rightarrow {}^{\max}\mathcal{Q}^i_{sa}(\mathbf{X})$, which is defined as

$$ {}^{\max}\mathcal{Q}^i_{sa}(\mathbf{X}) := \sum_{s'}\sum_{\mathbf{a}^{-i}} \mathbf{X}^{-i}_{s\mathbf{a}^{-i}} T_{s a \mathbf{a}^{-i} s'} \max_b Q^i_{s'b}(\mathbf{X}) \tag{16} $$

using the deterministic limit conversion rule [Eq. (15)]. Because of the assumption of infinite interactions, we can here replace the state-action-value approximations $\tilde{Q}^i_{s(t+1)b}$ with the true state-action values $Q^i_{s'b}$ as defined by Eq. (7).

For the third term, we invert Eq. (10), yielding $\tilde{Q}^i_{sa}(t) = (\beta^i)^{-1}\log X^i_{sa}(t) + \mathrm{const}^i_s$, where const^i_s is constant in actions but may vary for each agent and state. Now, one can show that the dynamics induced by Eq. (11) are invariant against additive transformations in the temporal difference error $D^i_{sa}(t,\infty) \rightarrow D^i_{sa}(t,\infty) + \mathrm{const}^i_s$. Thus, the third term can be converted according to $\tilde{Q}^i_{s(t)a(t)}(t) \rightarrow (\beta^i)^{-1}\log X^i_{sa}(t)$.

All together, the temporal difference error for Q learning in the deterministic limit reads

$$ \begin{aligned} {}^{q}D^i_{sa}(t,\infty) = {} & (1-\gamma^i)_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} \\ & + \gamma^{i\,\max}\mathcal{Q}^i_{sa}(\mathbf{X}) - \frac{1}{\beta^i}\log X^i_{sa}(t). \end{aligned} \tag{17} $$

2. SARSA learning

Two of the three terms of the SARSA temporal difference error are identical to the one of Q learning, leaving $\tilde{Q}^i_{s(t+1)a(t+1)}(t)$, which we replace by

$$ {}^{\mathrm{next}}\mathcal{Q}^i_{sa}(\mathbf{X}) := \sum_{s'}\sum_{\mathbf{a}^{-i}} \mathbf{X}^{-i}_{s\mathbf{a}^{-i}} T_{s a \mathbf{a}^{-i} s'} \sum_b X^i_{s'b} Q^i_{s'b}(\mathbf{X}) \tag{18} $$

using again the deterministic limit conversion rule [Eq. (15)] and the state-action value $Q^i_{s'b}(\mathbf{X})$ of the behavior profile $\mathbf{X}$ according to Eq. (7).

Thus, the temporal difference error for the SARSA learning update in the deterministic limit reads

$$ \begin{aligned} {}^{\mathrm{sarsa}}D^i_{sa}(t;\infty) = {} & (1-\gamma^i)_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} \\ & + \gamma^{i\,\mathrm{next}}\mathcal{Q}^i_{sa}(\mathbf{X}) - \frac{1}{\beta^i}\log X^i_{sa}(t). \end{aligned} \tag{19} $$

3. Actor-critic learning

For the temporal difference error for AC learning we have to find replacements for (i) $\tilde{V}^i_{s(t+1)}(t)$ and (ii) $\tilde{V}^i_{s(t)}(t)$. Applying again Eq. (15) yields $\tilde{V}^i_{s(t+1)}(t) \rightarrow {}^{\mathrm{next}}\mathcal{V}^i_{sa}$, defined as

$$ {}^{\mathrm{next}}\mathcal{V}^i_{sa} := \sum_{s'}\sum_{\mathbf{a}^{-i}} \mathbf{X}^{-i}_{s\mathbf{a}^{-i}} T_{s a \mathbf{a}^{-i} s'} V^i_{s'}(\mathbf{X}), \tag{20} $$

using Eq. (6) for the state value $V^i_{s'}(\mathbf{X})$. This is the average value of the next state given that in the current state the agent took action a. One can show that ${}^{\mathrm{next}}\mathcal{V}^i_{sa}(\mathbf{X}) = {}^{\mathrm{next}}\mathcal{Q}^i_{sa}(\mathbf{X})$ from the SARSA update.

The second remaining term belongs to the slower adaptation timescale or, in other words, occurs outside the batch. Thus, our deterministic limit conversion rule [Eq. (15)] does not apply. We could think of a conversion $\tilde{V}^i_{s(t)}(t) := \sum_a X^i_{sa}\tilde{Q}^i_{s(t)a(t)}(t) \rightarrow (\beta^i)^{-1}\sum_a X^i_{sa}(t)\log X^i_{sa}(t)$. However, the remaining term is constant in action, and therefore irrelevant for the dynamics, as we have argued above. Thus, we can simply put $\tilde{V}^i_{s(t)}(t) \rightarrow 0$.

All together, the temporal difference error of the actor-critic learner in the deterministic limit reads

$$ {}^{\mathrm{ac}}D^i_{sa}(t,\infty) = (1-\gamma^i)_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^{i\,\mathrm{next}}\mathcal{V}^i_{sa}(\mathbf{X}). \tag{21} $$

IV. APPLICATION TO EXAMPLE ENVIRONMENTS

In the following we apply the derived deterministic learning equations in two different environments. Specifically, we compare the three well-established temporal difference learning variants (Q learning, SARSA learning, and AC learning) in two different two-agent ($N = 2$), two-action ($M = 2$), and two-state ($Z = 2$) environments: a two-state matching pennies game and a two-state prisoner's dilemma. Since the main contribution of this paper is the derivation of the deterministic temporal difference learning equations, we are not trying to make a case with our example environments beyond a systematic comparison of our learners. Therefore, we chose environments that have been used previously in related literature [33–35,45]. Note also that we leave a comparison between the deterministic limit and the stochastic equations to future work, which would add a noise term to our equations following the example of Ref. [20].

To measure the performance of an agent's behavior profile in a single scalar, we use the dot product between the stationary state distribution $\boldsymbol{\sigma}(\mathbf{X})$ of the effective Markov Chain with the transition matrix ${}_{\mathbf{X}}\langle\mathbf{T}\rangle$ and the behavior average reward ${}_{\mathbf{TX}}\langle\mathbf{R}\rangle^i$. Interestingly, we find this relation to be identical to the dot product of the stationary distribution and the state value $\mathbf{V}^i(\mathbf{X})$:

$$ \boldsymbol{\sigma}(\mathbf{X}) \cdot {}_{\mathbf{TX}}\langle\mathbf{R}\rangle^i = \boldsymbol{\sigma}(\mathbf{X}) \cdot \mathbf{V}^i(\mathbf{X}). \tag{22} $$

This relation can be shown by using Eq. (6) and the fact that $\boldsymbol{\sigma}(\mathbf{X})$ is an eigenvector of ${}_{\mathbf{X}}\langle\mathbf{T}\rangle$.

In the following examples we will only investigate homogeneous agents, i.e., agents whose parameters will not differ from each other. We will therefore drop the agent indices from α^i, β^i, and γ^i. The heterogeneous agent case is to be explored in future work.

A. Two-state matching pennies

The single-state matching pennies game is a paradigmatic two-agent, two-action game. Imagine the situation of soccer

Reprinted paper with permission from W. Barfuss et al., Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E, vol. 99, no. 4, pp. 1-16, 2019, doi: 10.1103/PhysRevE.99.043305. Copyright 2019 by the American Physical Society

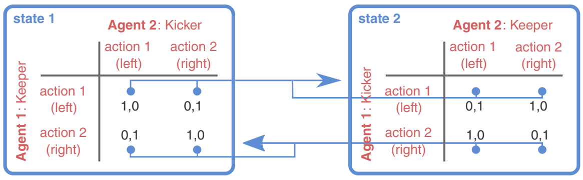


FIG. 2. Two-state matching pennies. Rewards are given in black type in the payoff tables for each state. State-transition probabilities are indicated by (blue) arrows.

penalty kicks. The keeper (agent 1) can choose to jump either to the left or right side of the goal, and the kicker (agent 2) can choose to kick the ball also either to the left or the right. If both agents choose the identical side, then the keeper agent wins; otherwise, the kicker wins.

In the two-state version of the game, according to Ref. [35], the rules are extended as follows: In state 1 the situation is as described in the single-state version. Whenever agent 1 (the keeper) decides to jump to the left, the environment transitions to state 2, in which the agents switch roles: Agent 1 now plays the kicker and agent 2 the keeper. From here, whenever agent 1 (now the kicker) decides to kick to the right side the environment transition again to state 1 and both agents switch their roles again.

Figure 2 illustrates this two-state matching pennies games. Formally, the payoff matrices are given by

$$\begin{pmatrix} R^1_{111s'}, R^2_{111s'} & R^1_{112s'}, R^2_{112s'} \\ R^1_{121s'}, R^2_{121s'} & R^1_{122s'}, R^2_{122s'} \end{pmatrix} = \begin{pmatrix} 1,0 & 0,1 \\ 0,1 & 1,0 \end{pmatrix}$$

in state 1 and

$$\begin{pmatrix} R^1_{211s'}, R^2_{211s'} & R^1_{212s'}, R^2_{212s'} \\ R^1_{221s'}, R^2_{221s'} & R^1_{222s'}, R^2_{222s'} \end{pmatrix} = \begin{pmatrix} 0,1 & 1,0 \\ 1,0 & 0,1 \end{pmatrix}$$

in state 2 for $s' \in \{1, 2\}$. State transitions are governed by

$$\begin{pmatrix} T_{1112} & T_{1122} \\ T_{1212} & T_{1222} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} T_{2111} & T_{2121} \\ T_{2211} & T_{2221} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix}.$$

Thus, by construction, the probability of transitioning to the other state is independent of agent 2's action. Only agent 1 has agency over the state transitions. By playing a uniformly random behavior profile $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.5, 0.5, 0.5, 0.5)$, both agents would obtain an average reward of 0.5 per time step.

With Fig. 3 we compare the temporal difference error in the behavior space sections for each environmental state at a comparable low discount factor $\gamma \in [0, 1)$ of $\gamma = 0.1$, as well as learning trajectories for an exemplary initial condition for two learning rates $\alpha \in (0, 1)$, a low one ($\alpha = 0.02$) and a high one ($\alpha = 0.8$). Overall, we observe a variety of qualitatively different dynamical regimes, such as fixed points, periodic orbits and chaotic motion.

Specifically, we see that Q learners and SARSA learners behave qualitatively similarly in contrast to the AC learners for both learning rates α. For the low learning rate $\alpha = 0.02$, Q and SARSA learners reach a fixed point of playing both actions with equal probability in both states, yielding a reward of 0.5. Due to the low α, this takes approximately 600 time steps. In contrast, the reward trajectory of the AC learners appears to be chaotic. Figure 5 confirms this observation, which we will discuss in more detail below.

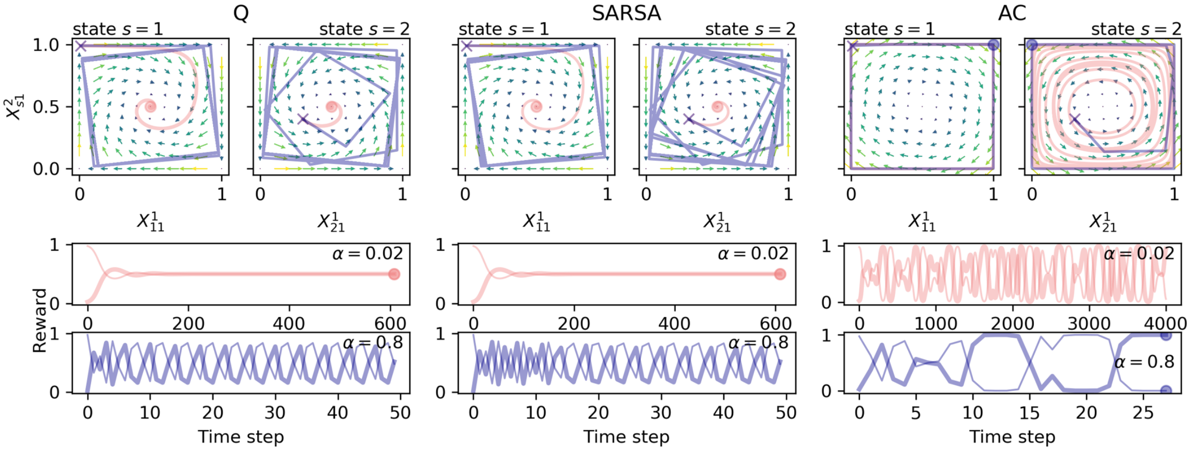


FIG. 3. Three learners in two-state matching pennies environment for low discount factor $\gamma = 0.1$; intensity of choice $\beta = 5.0$. At the top, the temporal difference errors for the Q learners [Eq. (17)], SARSA learners [Eq. (19)], and AC learners [Eq. (21)] are shown in two behavior phase-space sections, one for each state. The arrows indicate the average direction the temporal difference errors drive the learner toward, averaged over all phase-space points of the other state. Arrow colors (and shadings) additionally encode their lengths. Selected trajectories are shown in the phase-space sections, as well as by reward trajectories, plotting the average reward value [Eq. (22)] over time steps. Crosses in the phase-space subsections indicate the initial behavior $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.01, 0.99, 0.3, 0.4)$. Circles signal the arrival at a fixed point, determined by the absolute difference of behavior profiles between two subsequent time steps being below $\epsilon = 10^{-6}$. Trajectories are shown for two different learning rates $\alpha = 0.02$ (light red) and $\alpha = 0.8$ (dark blue). The bold reward trajectory belongs to agent 1 and the thin one to agent 2. Note that the temporal difference error is independent from the learning rate α. A variety of qualitatively different dynamical regimes can be observed.

Reprinted paper with permission from W. Barfuss et al., Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E, vol. 99, no. 4, pp. 1-16, 2019, doi: 10.1103/PhysRevE.99.043305. Copyright 2019 by the American Physical Society

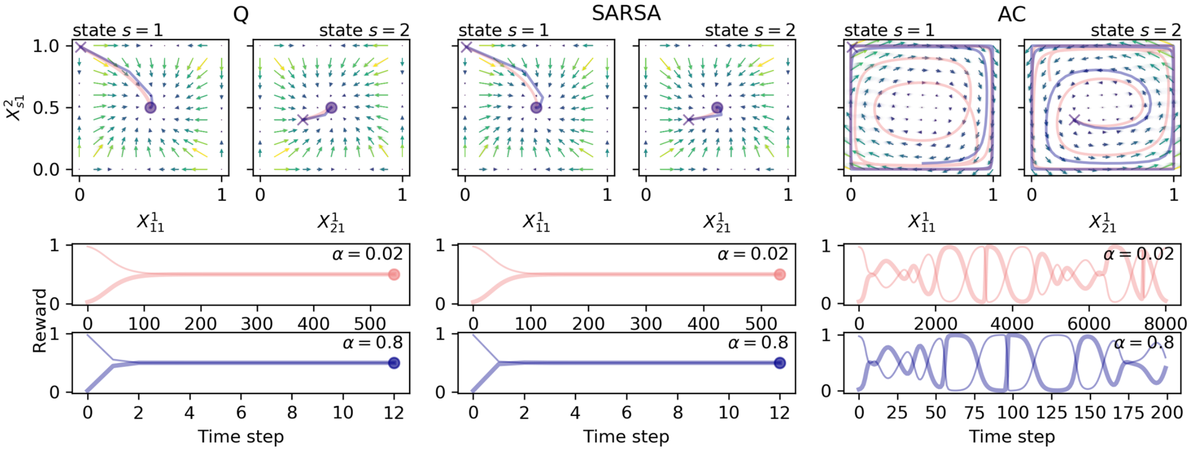


FIG. 4. Two-state matching pennies environment for high discount factor $\gamma = 0.9$; otherwise, identical to Fig. 3.

For the high learning rate $\alpha = 0.8$, both Q and SARSA learners enter a periodic limit cycle. Differences in the trajectories of Q and SARSA learners are clearly visible. The time average reward of this periodic orbit appears to be approximately 0.5 for each agent, identical to the reward of the fixed point at lower α. The AC learners, however, converge to a fixed point after oscillating near the edges of the phase space. At this fixed point agent 1 plays action 1 in state 1 with probability 1. Thus, it has trapped the system into state 2. In state 2, agent 1 plays action 2 and agent 2 plays action 1 with probability 1 and, consequently, agent 1 receives a reward of 1, whereas agent 2 receives 0 reward. One might ask, Why does agent 2 not decrease her probability for playing action 1, thereby increasing her own reward? And, indeed, the arrows of the temporal difference error suggest this change of behavior profile. However, agent 2 cannot follow because her behavior is trapped on the simplex of nonzero action probabilities X^2_{2a}. For only $M = 2$ actions, $X^2_{21} = 1$ thus can no longer change, regardless of the temporal difference error.

Increasing the discount factor to $\gamma = 0.9$, we observe the learning rate α to set the timescale of learning (Fig. 4). The intensity of choice remained $\beta = 5.0$. A high learning rate $\alpha = 0.8$ corresponds to faster learning in contrast to a low learning rate $\alpha = 0.02$. Also, the ratio of learning timescales is comparable to the inverse ratio of learning rates. For both α, Q and SARSA learners reach a fixed point, whereas the AC learners seem to move chaotically (details to be investigated below). Comparing the trajectories between the learning rates α, we observe a similar shape for each pair of learners. However, the similarity of the AC trajectories decreases at larger time steps.

So far, we varied two parameters: the discount factor $\gamma \in [0, 1)$ and the learning rate $\alpha \in (0, 1)$. Combining Figs. 3 and 4, we investigated all four combinations of a low and a high γ with a low and a high α. We can summarize that Q and SARSA learners converge to a fixed point for all combinations of discount factor γ and learning rate α, except when γ is low and α simultaneously high. Actor-critic dynamics seem chaotic for all combinations of α and γ.

To investigate the relationship between the parameters more thoroughly, Fig. 5 shows bifurcation diagrams with the bifurcation parameters α and γ. Additionally, it also gives the largest Lyapunov exponents for each learner and each parameter combination. A largest Lyapunov exponent greater than zero is a key characteristic of chaotic motion. We computed the Lyapunov exponent from the analytically derived Jacobian matrix, iteratively used in a QR decomposition according to Ref. [46]. See Appendix for details.

The largest Lyapunov exponent for Q and SARSA learners align almost perfectly with each other, whereas the largest Lyapunov exponent of the AC learners behaves qualitatively different. We first describe the behavior of the Q and SARSA learner: For high learning rates α and low farsightedness γ, Fig. 5 shows a periodic orbit with few (four) points in phase space. Largest Lyapunov exponents are distinctly below 0 at those regimes. Increasing the farsightedness γ both learners enter a regime of visiting many points in phase space around the stable fixed point $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.5, 0.5, 0.5, 0.5)$. The largest Lyapuonv exponents are close to zero. With increasing γ the distance around this fixed-point solution decreases until the dynamics converge from a farsightedness γ slightly greater than 0.5 onward. From there the largest Lyapunov exponent decreases again for further increasing γ. The same observations can be made along a decreasing bifurcation parameter α, except that at the end, for low α, the largest Lyapunov exponents do not decrease as distinctly as for high γ.

The behavior of the actor-critic dynamics is qualitatively different from the one of Q and SARSA. The placement of the fixed points on the natural numbers grid suggests that the AC learners get confined on one of the 16 (M^{NZ}) corners of the behavior phase space. No regularity to which fixed point the AC learners converge can be deduced. The largest Lyapunov exponent is always above zero and experiences an overall decreasing behavior. Similarly, for a decreasing bifurcation parameter α, the largest Lyapunov exponent tends to decrease as well. Different from the bifurcation diagram along γ, for low α the system might enter a periodic motion but only

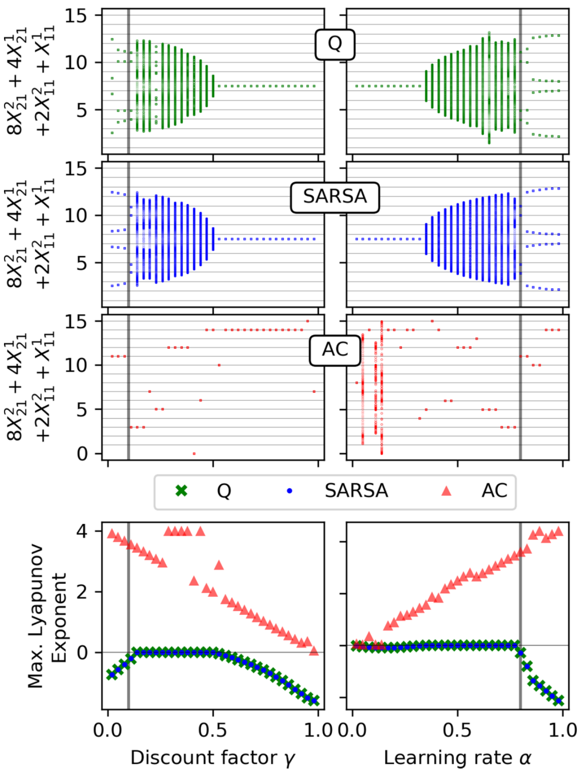


FIG. 5. Varying discount factor γ and learning rate α in two-state matching pennies environment for intensity of choice $\beta = 5.0$ for the Q learners (green crosses), the SARSA learners (blue dots), and the AC learners (red triangles). On the left, the discount factor γ is varied with learning rate $\alpha = 0.8$, as indicated by the gray vertical lines on the right. On the right, the learning rate α is varied with discount factor $\gamma = 0.1$ as indicated by the gray vertical lines on the left. The three top panels show the visited behavior points during 1000 iterations after a transient period of 100 000 time steps from initial behavior $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.01, 0.99, 0.3, 0.4)$. Visited points are mapped to the function $8X^2_{21} + 4X^1_{21} + 2X^2_{11} + X^1_{11}$ on the vertical axes to give a fuller image of the visited behavior profiles. The bottom panel shows the corresponding largest Lyapunov exponents for the three learners. Overall, Q and SARSA learners behave qualitatively more similarly than the actor-critic learners.

for some parameters α. No regularity can be determined at which parameters α the AC learners enter a periodic motion. A more thorough investigation of the nonlinear dynamics, especially those of the actor-critic learner, seems of great interest but is, however, beyond the scope of this article and leaves promising paths for future work.

Concerning the parameter β, the intensity of choice, one can infer from the update equations [Eq. (11) combined with Eq. (19) and Eq. (21)] that the dynamics for the AC learners are invariant for a constant product $\alpha\beta$. This is because the temporal difference error of the actor-critic learners in the deterministic limit is independent of β. Further, the dynamics of the SARSA learners will converge to the dynamics of the AC learners under $\beta \to \infty$. Figure 6 nicely confirms these two observations. Observing Table I is another way to see this. Since the value estimate of the future state is identical for SARSA and AC learning, letting the value estimate of the current state vanish by sending $\beta \to \infty$ makes the SARSA learners approximate the AC learners.

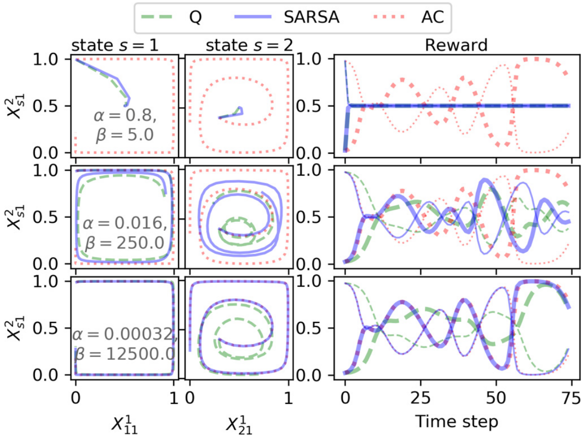


FIG. 6. Varying intensity of choice β under constant $\alpha\beta$ in a two-state matching pennies environment for discount factor $\gamma = 0.9$. On the left trajectories of the three learners [Q, green dashed; SARSA, blue straight; AC, red dotted] are shown in the two behavior space sections, one for each state. On the right, the corresponding reward trajectories are shown. The initial behavior was $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.01, 0.99, 0.3, 0.4)$. The bold reward trajectory belongs to agent 1 and the thin one to agent 2. One observes the deterministic limit of actor-critic learning to be invariant under constant $\alpha\beta$ and SARSA learning to converge to AC learning under $\beta \to \infty$.

As mentioned before, β controls the exploration-exploitation trade-off. In the temporal difference errors of Q and SARSA learning it appears in the term indicating the value estimate of the current state $-1/\beta^i \log(X^i_{sa})$. If this term dominates the temporal difference error (i.e., if β is small), then the learners tend toward the center of behavior space, i.e., $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.5, 0.5, 0.5, 0.5)$, forgetting what they have learned about the obtainable reward. This characteristic happens to be favorable in our two-state matching pennies environment, which is why Q and SARSA learners perform better in finding the $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.5, 0.5, 0.5, 0.5)$ solution. On the other hand, if β is large, then the temporal difference error is dominated by the current reward and future value estimate. Not being able to forget, the learners might get trapped in unfavorable behavior, as we can see observing the actor-critic learners. To calibrate β it is useful to make oneself clear that it must come in units of [log behavior]/[reward].

B. Two-state prisoner's dilemma

The single-state prisoner's dilemma is another paradigmatic two-agent, two-action game. It has been used to model social dilemmas and study the emergence of cooperation. It describes a situation in which two prisoners are separately interrogated, leaving them with the choice to either cooperate with each other by not speaking to the police or defecting by testifying.

Reprinted paper with permission from W. Barfuss et al., Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E, vol. 99, no. 4, pp. 1-16, 2019, doi: 10.1103/PhysRevE.99.043305. Copyright 2019 by the American Physical Society

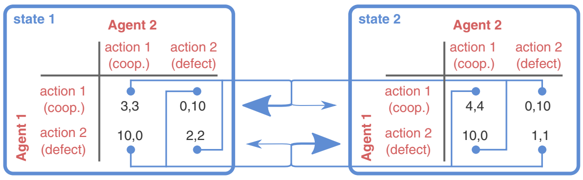


FIG. 7. Two-state prisoner's dilemma. Rewards are given in black type in the payoff tables for each state. State-transition probabilities are indicated by (blue) arrows.

The two-state version, which has been used as a test environment also in Refs. [33–35], extends this situation somewhat artificially by playing a prisoner's dilemma in each of the two states with a transition probability of 10% from one state to the other if both agents chose the same action and a transition probability of 90% if both agents chose opposite actions.

Figure 7 illustrates these game dynamics. Formally, the payoff matrices are given by

$$\begin{pmatrix} R^1_{111s'}, R^2_{111s'} & R^1_{112s'}, R^2_{112s'} \\ R^1_{121s'}, R^2_{121s'} & R^1_{122s'}, R^2_{122s'} \end{pmatrix} = \begin{pmatrix} 3, 3 & 0, 10 \\ 10, 0 & 2, 2 \end{pmatrix}$$

in state 1 and

$$\begin{pmatrix} R^1_{211s'}, R^2_{211s'} & R^1_{212s'}, R^2_{212s'} \\ R^1_{221s'}, R^2_{221s'} & R^1_{222s'}, R^2_{222s'} \end{pmatrix} = \begin{pmatrix} 4, 4 & 0, 10 \\ 10, 0 & 1, 1 \end{pmatrix}$$

in state 2 for $s' \in \{1, 2\}$, respectively. The corresponding state transition probabilities are given by

$$\begin{pmatrix} T_{1112} & T_{1122} \\ T_{1212} & T_{1222} \end{pmatrix} = \begin{pmatrix} T_{2111} & T_{2121} \\ T_{2211} & T_{2221} \end{pmatrix} = \begin{pmatrix} 0.1 & 0.9 \\ 0.9 & 0.1 \end{pmatrix}.$$

To be precise, the rewards in each state do not resemble a classical social dilemma situation. This is because if both agents would alternately cooperate and defect, both could receive a larger reward per time step compared to always cooperating. Hence, this stochastic game, as it was used in Refs. [33–35], presents more a coordination than a cooperation challenge to the agents. The multistate environment can here function as a coordination device.

A behavior profile in which one agent exploits the other in one state, while being exploited in the other state, would result in an average reward per time step of 5 for each agent, e.g., $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0, 1, 1, 0)$.

However, for all three learning types with a midranged farsightedness ($\gamma = 0.45$) and an intensity of choice $\beta = 5.0$, the temporal difference error arrows are pointing on average toward the lower-left defection-defection point for each state in behavior phase space (Fig. 8). To see whether the three learning types may converge to the described defect-cooperate–cooperate-defect equilibrium, individual trajectories from two exemplary initial conditions and for two learning rates α are shown, a small one ($\alpha = 0.02$) and a high one ($\alpha = 0.8$).

We observe qualitatively different behavior across all three learners. The Q learners converge to equilibria with average rewards distinctly below 5, and the SARSA learners converge to equilibira with average rewards of almost 5 for both learning rates α and both exemplary initial conditions. Both Q and SARSA learners converge to solutions of proper probabilistic behavior, i.e., choosing action cooperate and action defect with nonvanishing chance. The actor-critic learners, on the other hand, converge to the deterministic defect-cooperate–cooperate-defect behavior described above for the initial condition shown with the nondashed lines in Fig. 8 for both learning rates α (shown in light red and dark blue). For the other exemplary initial condition, shown with the dashed lines, it converges to an all-defection solution in both states for both α.

Interestingly, for all learners, all combinations of initial conditions and learning rates converge to a fixed point solution, except for the Q learners with a comparably high learning rate $\alpha = 0.8$, which enter a periodic behavior solution for the initial condition with the nondashed line. The same phenomenon occurred also in the matching pennies environment for low farsightedness $\gamma = 0.1$, however, there for both Q and SARSA learners. It seems to be caused by the comparably high learning rate. A high learning rate overshoots the behavior update, resulting in a circling behavior

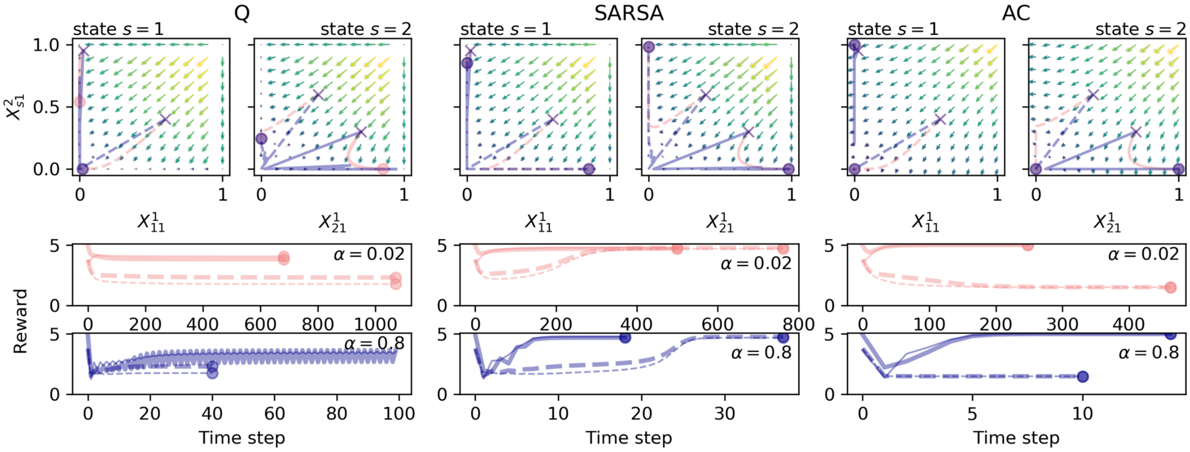


FIG. 8. Two-state prisoner's dilemma environment for discount factor $\gamma = 0.45$; otherwise, identical to Fig. 3.

Reprinted paper with permission from W. Barfuss et al., Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E, vol. 99, no. 4, pp. 1-16, 2019, doi: 10.1103/PhysRevE.99.043305. Copyright 2019 by the American Physical Society

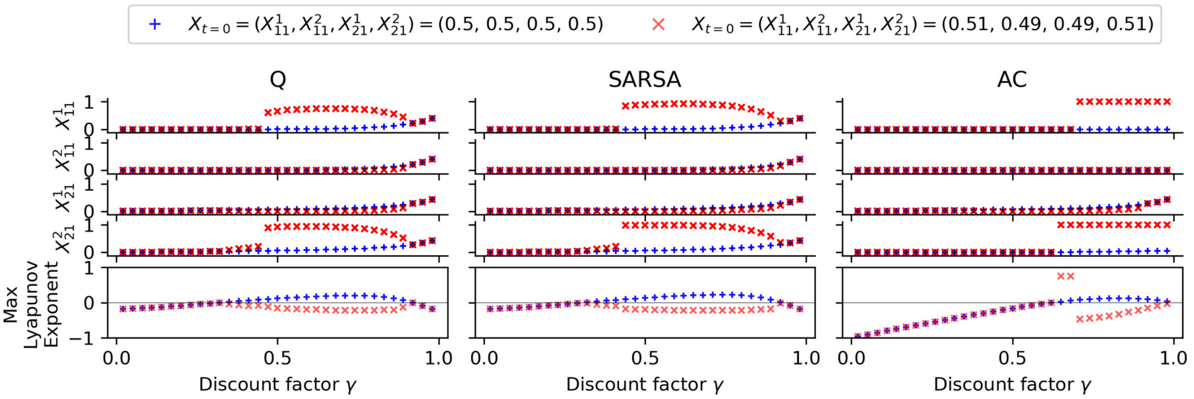


FIG. 9. Varying discount factor γ in two-state prisoner's dilemma environment for learning rate $\alpha = 0.2$ and intensity of choice $\beta = 5.0$ for the Q learners on the left, the SARSA learners in the middle, and the actor-critic learners on the right. The four top panels for each learner show the visited behavior points $X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}$ during 1000 iterations after a transient period of 5000 time steps from initial behavior $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.5, 0.5, 0.5, 0.5)$ in blue pluses and from initial behavior $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.51, 0.49, 0.49, 0.51)$ in red crosses. The bottom panels show the corresponding largest Lyapunov exponents for the two initial conditions. Above a critical discount factor γ all learners find the high rewarding solution from the red crosses initial condition but do not do so from the blue pluses initial condition.

around the fixed point. As in Fig. 3, the time average reward of the periodic orbit seems to be comparable to the reward of the corresponding fixed point at lower α. Furthermore, we observe the same time rescaling effect of the learning rate α in Fig. 8 as in Fig. 4.

To visualize the influence of the discount factor γ on the converged behavior, Fig. 9 shows a bifurcation diagram along the bifurcation parameter γ for two initial conditions. Pluses in blue result from a uniformly random behavior profile of $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.5, 0.5, 0.5, 0.5)$, whereas the crosses in red initially started from the behavior profile $(X^1_{11}, X^2_{11}, X^1_{21}, X^2_{21}) = (0.51, 0.49, 0.49, 0.51)$.

Across all learners, lower discount factors γ correspond to all-defect solutions, whereas for higher γ the solutions from the initial condition shown with red crosses tend toward the cooperate-defect–defect-cooperate solution. For low γ, the agents are less aware of the presence of other states and find the all-defect equilibrium solution of the iterated normal form prisoner's dilemma. The state transition probabilities have less effect on the learning dynamics. Only above a certain farsightedness do the agents find the more rewarding cooperate-defect–defect-cooperate solution.

The observation from Fig. 8 is confirmed that the probability to cooperate (i.e., here X^1_{11} and X^2_{21}) is lowest for the Q learners, midrange for the SARSA learners, and 1 for the actor-critic learners. One reason for this observation can be found in the intensity of choice parameter β. It balances the reward obtainable in the current behavior space segment with the forgetting of current knowledge to be open to new solutions. Such forgetting expresses itself by temporal difference error components pointing toward the center of behavior space. Thus, a relatively small $\beta = 5.0$ can explain why solutions at the edge of the behavior space cannot be reached by Q and SARSA learners. The AC learners mis this forgetting term in the deterministic limit and can therefore easily enter behavior profiles at the edge of the behavior space.

Q and SARSA learners have a critical discount factor γ above which the cooperate-defect–defect-cooperate high reward solution is obtained and below which the all-defect low reward solution gets selected. However, for increasing discount factors γ up to 1, Q and SARSA learners experience a drop in playing the cooperative action probability.

The actor-critic learners approach the cooperate-defect–defect-cooperate solution in two steps. For increasing γ, first the cooperation probability of agent 2 in state 2 (X^2_{21}) jumps from zero to 1 while agent 1 still defects in state 1. Only after a slight increase of γ does agent 1 then also cooperate in state 1 (X^1_{11}).

Interestingly, for the uniformly random initial behavior condition shown with blue pluses, there is no critical discount factor γ and no learners come close to the cooperate-defect–defect-cooperate solution. Here, only for γ close to 1 do all cooperation probabilities X^i_{s1} gradually increase. Furthermore, exactly at those γ, where the cooperate-defect–defect-cooperate solution is obtained from the initial behavior condition shown with red crosses, the solutions from the uniformly random initial behavior condition (blue pluses) have a largest Lyapuonv exponent greater than 0. At other values of γ, the largest Lyapunov exponents for the two initial conditions overlap. This suggests that the largest Lyapunov exponents greater than zero may point to the fact that other, perhaps more rewarding, solutions may exist in phase space. A more thorough investigation regarding this multistability is an open point for future research.

As we have argued above, the two-state prisoner's dilemma as it was used in Refs. [33–35] presents rather a coordination than a cooperation challenge to the agents. Figure 10 demonstrates that our learning dynamics are also capable of solving a cooperation challenge in a stochastic game setting, for which we adapt a two-state prisoner's dilemma in analogy to Ref. [45]. Figure 10 confirms previous findings that cooperation emerges only in the stochastic game, compared to

Reprinted paper with permission from W. Barfuss et al., Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E, vol. 99, no. 4, pp. 1-16, 2019, doi: 10.1103/PhysRevE.99.043305. Copyright 2019 by the American Physical Society

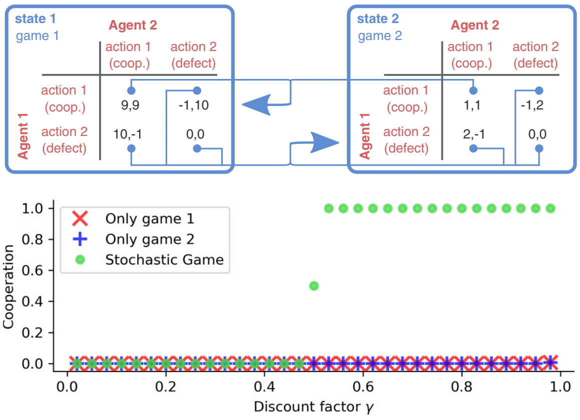


FIG. 10. Cooperation challenge in a two-state prisoner's dilemma. Top panel shows a two-state prisoner's dilemma game, whose state games individually favor defection. Bottom panel shows the level of cooperation SARSA learners with $\alpha = 0.016$, $\beta = 250$ play after reaching a fixed point from the center of behavior space ($X^i_{sa} = 0.5$ for all i, s, a) for varying discount factors γ. Results for Q and AC learners are similar. Cooperation levels are shown for the full stochastic game as well as for each individual state game played repeatedly. For sufficiently large farsightedness, cooperation can emerge in the stochastic game, in contrast to the individual repeated games.

playing each prisoner's dilemma repeatedly [45]. Further, cooperation only emerges for sufficiently large farsightedness γ.

V. DISCUSSION

The main contribution of this paper is the development of a technique to obtain the deterministic limit of temporal difference reinforcement learning. Through our work we have combined the literature on learning dynamics from statistical physics with the evolutionary game theory-inspired learning dynamics literature from machine learning. For the statistical physics community, we present learning equations capable of handling environmental state transitions. For the machine-learning community, we present the systematic methodology we have used to obtain the deterministic learning equations.

We have demonstrated our approach with the three prominent reinforcement learning algorithms from computer science: Q learning, SARSA learning, and actor-critic learning. A comparison of their dynamics in previously used two-agent, two-action, two-state environments has revealed the existence of a variety of qualitatively different dynamical regimes, such as convergence to fixed points, periodic orbits, and deterministic chaos.

We have found that Q and SARSA learners tend to behave qualitatively more similar in comparison to the actor-critic learning dynamics. This characteristic results at least partly from our relatively low intensity of choice parameter β, controlling the exploration-exploitation trade-off via a forgetting term in the temporal difference errors. Sending $\beta \rightarrow \infty$, the SARSA learning dynamics approach the actor-critic learning dynamics, as we have shown. Overall the actor-critic learners have a tendency to enter confining behavior profiles, due to their nonexisting forgetting term. This characteristic leaves them trapped at the edges of the behavior space. In contrast, Q and SARSA learner do not show such learning behavior. Interestingly, this characteristic of the AC learners turns out to be favorable in the two-state prisoner's dilemma environment, where they find the most rewarding solution in more cases compared to Q and SARSA but hinders the convergence to the fixed point solution in the two-state matching pennies environment. Thus, the most favorable level of forgetting depends on the environment. In order to tune the respective parameter β, our consideration that it must come in the unit of [log behavior]/[reward] may be helpful.

We have demonstrated the effect of the learning rate α adjusting the speed of learning by controlling the amount of new information used in a behavior profile update. Thereby, within limits, α functions as a time rescaling. However, a comparably large learning rate α might cause an overshooting phenomenon, hindering the convergence to a fixed point. Instead, the learners enter a limit cycle around that point. Nevertheless, the average reward of the limit-cycling behavior was approximately equal to the one of the fixed point obtained at lower α but took fewer time steps to reach. Thus, perhaps other dynamical regimes than fixed points, such as limit cycles or strange attractors, could be of interest in some applications of reinforcement learning.

We have also shown the effect of the discount factor γ adjusting the farsightedness of the agents. At low γ the state transition probabilities have less effect on the learning dynamics compared to high discount factors.

To summarize the three parameters α, β, and γ: The level of exploitation β and the farsightedness γ control *where* the learners adapt toward in behavior space, weighting current reward, expected future reward, and the level of forgetting. The learning rate α controls *how fast* the learners adapt along these directions.

We hope that our work might turn out useful for the application of reinforcement learning in various domains, with respect to parameter tuning, the design of new algorithms, and the analysis of complex strategic interactions using meta strategies, as Bloembergen *et al.* [28] have pointed out. In this regard, future work could extend the presented methodology to partial observability of the Markov states of the environment [40,41], behavior profiles with history, and other-regarding agent (i.e., joint-action) learners (cf. Ref. [2] for an overview of other-regarding agent learning algorithms). Also, the combination of individual reinforcement learning and social learning through imitation [47–50] seems promising. Such endeavors would naturally lead to the exploration of network effects. It is important to note that only a few dynamical systems reinforcement learning studies have begun to incorporate network structures between agents [22,23].

Apart from these more technical extensions, we hope that our learning equations will prove themselves useful when studying the evolution of cooperation in stochastic games [45]. With stochastic games one is able to explicitly account for a changing environment. Therefore, such studies are likely to contribute to the advancement of theoretical research on the sustainability of interlinked social-ecological systems [51,52]. Interactions, synergies, and trade-offs between social [13,53] and ecological [54] dilemmas

Reprinted paper with permission from W. Barfuss et al., Deterministic limit of temporal difference reinforcement learning for stochastic games, Phys. Rev. E, vol. 99, no. 4, pp. 1-16, 2019, doi: 10.1103/PhysRevE.99.043305. Copyright 2019 by the American Physical Society

can be explored using the framework of stochastic games. More realistic environments, modeling, e.g., the harvesting of common-pool renewable resources [55,56] or the prevention of dangerous climate change [57,58], for our learning dynamics are likely to prove themselves useful. Here, it may be of interest to evaluate the learning process not only in terms of efficiency but also how close it came to the optimal behavior. Other paradigms than value optimization may also be important [59], such as sustainability or resilience [60].

Python code for reproduction of the figures of this article is available online at [61].

ACKNOWLEDGMENTS

This work was developed in the context of the COPAN project on Coevolutionary Pathways in the Earth System at the Potsdam Institute for Climate Impact Research (PIK). We are grateful for financial support by the Heinrich Böll Foundation, the Stordalen Foundation (via the Planetary Boundaries Research Network PB.net), the Earth League's EarthDoc program, and the Leibniz Association (project DOMINOES). We thank Jobst Heitzig for discussions and comments on the manuscript.

APPENDIX: COMPUTATION OF LYAPUNOV EXPONENTS

We compute the Lyapunov exponents using an iterative QR decomposition of the Jacobian matrix according to Sandri [46]. In the following we present the derivation of the Jacobian matrix.

Equation (11) constitutes a map f, which iteratively updates the behavior profile $\mathbf{X} \in \mathbb{R}^{N\times M\times Z}$. Consequently, we can represent its derivative as a Jacobian tensor $f'(\mathbf{X}) \in \mathbb{R}^{N\times M\times Z\times N\times M\times Z}$.

Let $A^i_{sa} := X^i_{sa} \exp[\alpha^i\beta^i D^i_{sa}(\mathbf{X})]$ be the numerator of Eq. (11) and $B^i_s := \sum_b A^i_{sb}$ its denominator, i.e., $f =: A/B$. Hence,

$$f'(\mathbf{X}) = \frac{A'B - B'A}{B^2} \tag{A1}$$

or, more precisely, in components,

$$\frac{df^i_{sa}(\mathbf{X})}{dX^j_{rb}} = \frac{\frac{dA^i_{sa}(\mathbf{X})}{dX^j_{rb}} B^i_s(\mathbf{X}) - \frac{dB^i_s}{dX^j_{rb}}(\mathbf{X}) A^i_{sa}(\mathbf{X})}{[B^i_s(\mathbf{X})]^2}. \tag{A2}$$

A and B are known, and if A' is known, then B' is easily obtained by $\frac{dB^i_s(\mathbf{X})}{dX^j_{rb}} = \sum_c \frac{dA^i_{sc}(\mathbf{X})}{dX^j_{rb}}$. Therefore we need to compute A' for the three learner types Q, SARSA, and actor-critic learning.

1. Q learning

Let us rewrite A^i_{sa} for the Q learner according to

$$A^i_{sa} := \left(X^i_{sa}\right)^{(1-\alpha^i)} \exp\left[\alpha^i\beta^i\hat{D}^i_{sa}(\mathbf{X})\right], \tag{A3}$$

where we removed the estimate of the current value from the temporal difference error, leaving the truncated temporal difference error as

$$\hat{D}^i_{sa}(\mathbf{X}) := (1-\gamma^i)\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^i\,{}^{\max}Q^i_{sa}(\mathbf{X}). \tag{A4}$$

Hence, we can write the derivative of A as

$$\frac{dA^i_{sa}(\mathbf{X})}{dX^j_{rb}} = \exp\left[\alpha^i\beta^i\hat{D}^i_{sa}(\mathbf{X})\right]\Bigg[(1-\alpha^i)\left(X^i_{sa}\right)^{-\alpha^i}\frac{dX^i_{sa}}{dX^j_{rb}} + \alpha^i\beta^i\left(X^i_{sa}\right)^{(1-\alpha^i)}\frac{d\hat{D}^i_{sa}(\mathbf{X})}{dX^j_{rb}}\Bigg]. \tag{A5}$$

Since $\sum_c X^i_{sc} = 1$, dX^i_{sa}/dX^j_{rb} can be expressed as

$$\frac{dX^i_{sa}}{dX^j_{rb}} = \delta_{ij}\delta_{sr}(2\delta_{ab} - 1). \tag{A6}$$

The derivative of the truncated temporal difference error reads

$$\frac{d\hat{D}^i_{sa}(\mathbf{X})}{dX^j_{rb}} = (1-\gamma^i)\frac{d\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}}{dX^j_{rb}} + \gamma^i\frac{d\,{}^{\max}Q^i_{sa}(\mathbf{X})}{dX^j_{rb}}. \tag{A7}$$

Let us write the derivative of the reward as

$$\frac{d\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}}{dX^j_{rb}} = \sum_{s'}\sum_{\mathbf{a}^{-i}}\frac{d\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}}{dX^j_{rb}}T_{sa\mathbf{a}^{-i}s'}R^i_{sa\mathbf{a}^{-i}s'} \tag{A8}$$

using Eq. (2) and Eq. (3), where the derivatives $d\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}/dX^j_{rb}$ need to be executed according to Eq. (A6).

For the derivative of the maximum next value we write accordingly

$$\frac{d\,{}^{\max}Q^i_{sa}(\mathbf{X})}{dX^j_{rb}} = \sum_{s'}\sum_{\mathbf{a}^{-i}}\frac{d\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}}{dX^j_{rb}}T_{sa\mathbf{a}^{-i}s'}\max_c Q^i_{s'c}(\mathbf{X}) + \sum_{s'}\sum_{\mathbf{a}^{-i}}\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}T_{sa\mathbf{a}^{-i}s'}\frac{d\max_c Q^i_{s'c}(\mathbf{X})}{dX^j_{rb}}. \tag{A9}$$

Let $a^m := \arg\max_a Q^i_{sa}(\mathbf{X})$, then

$$\frac{d\max_c Q^i_{sc}(\mathbf{X})}{dX^j_{rb}} = \delta_{aa^m}\frac{dQ^i_{sa}(\mathbf{X})}{dX^j_{rb}} \tag{A10}$$

and

$$\frac{dQ^i_{sa}(\mathbf{X})}{dX^j_{rb}} = (1-\gamma^i)\frac{d\,{}_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}}{dX^j_{rb}} + \gamma^i\sum_{s'}\frac{d\,{}_{\mathbf{X}}\langle T\rangle_{ss'}}{dX^j_{rb}}V^i_{s'}(\mathbf{X}) + {}_{\mathbf{X}}\langle T\rangle_{ss'}\frac{dV^i_{s'}(\mathbf{X})}{dX^j_{rb}}. \tag{A11}$$

For the derivative of the effective Markov Chain transition tensor we can write

$$\frac{d\,{}_{\mathbf{X}}\langle T\rangle_{ss'}}{dX^j_{rb}} = \sum_{\mathbf{a}}\frac{d\mathbf{X}_{s\mathbf{a}}}{dX^j_{rb}}T_{sa\mathbf{a}^{-i}s'}, \tag{A12}$$

using Eqs. (2) and (3), where again the derivatives $d\mathbf{X}_{s\mathbf{a}}/dX^j_{rb}$ need to be executed according to Eq. (A6).

For the derivative of the state value let us rewrite Eq. (6) as $V^i_s = (1-\gamma^i)\sum_{s'} M^{-1}_{ss'}\,{}_{\mathbf{TX}}\langle R\rangle^i_{s'}$ with $M := (\mathbb{1}_Z - \gamma^i{}_{\mathbf{X}}\langle T\rangle)$. Thus,

$$\frac{dV^i_s(\mathbf{X})}{dX^j_{rb}} = (1-\gamma^i)\sum_{s''}\frac{d\big(M^{-1}_{ss''}\big)}{dX^j_{rb}}\,{}_{\mathbf{TX}}\langle R\rangle^i_{s''} + M^{-1}_{ss''}\frac{d_{\mathbf{TX}}\langle R\rangle^i_{s''}}{dX^j_{rb}}. \tag{A13}$$

To obtain the derivative of the inverse matrix M^{-1} we use $(M^{-1}M)' = 0 = (M^{-1})'M + M^{-1}M'$ and therefore $(M^{-1})' = -M^{-1}M'M^{-1}$. For M' we write

$$\frac{dM_{ss'}}{dX^j_{rb}} = -\gamma^i\frac{d_{\mathbf{X}}\langle T\rangle_{ss'}}{dX^j_{rb}}. \tag{A14}$$

We obtain the derivative of the reward according to

$$\frac{d_{\mathbf{TX}}\langle R\rangle^i_s}{dX^j_{rb}} = \sum_{s'}\sum_{\mathbf{a}}\frac{d\mathbf{X}_{s\mathbf{a}}}{dX^j_{rb}}T_{s\mathbf{a}s'}R^i_{s\mathbf{a}s'}, \tag{A15}$$

using Eq. (1) and Eq. (3), where the derivatives dX^i_{sa}/dX^j_{rb} need to be executed according to Eq. (A6).

Now we can compute the Jacobian matrix for the Q learning dynamics in their deterministic limit.

2. SARSA learning

The computation of the Jacobian matrix for the SARSA learning update in its deterministic limit is similar, except the truncated temporal difference error reads

$$\hat{D}^i_{sa}(\mathbf{X}) := (1-\gamma^i)_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^{i\,\text{next}}\mathcal{Q}^i_{sa}(\mathbf{X}) \tag{A16}$$

instead of Eq. (A4). Hence,

$$\frac{d\hat{D}^i_{sa}(\mathbf{X})}{dX^j_{rb}} = (1-\gamma^i)\frac{d_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa}}{dX^j_{rb}} + \gamma^i\frac{d^{\text{next}}\mathcal{Q}^i_{sa}(\mathbf{X})}{dX^j_{rb}} \tag{A17}$$

and

$$\frac{d^{\text{next}}\mathcal{Q}^i_{sa}(\mathbf{X})}{dX^j_{rb}} = \sum_{s'}\sum_{\mathbf{a}^{-i}}\frac{d\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}}{dX^j_{rb}}T_{sa\mathbf{a}^{-i}s'}\sum_c X^i_{s'c}\mathcal{Q}^i_{s'c}(\mathbf{X}) + \sum_{s'}\sum_{\mathbf{a}^{-i}}\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}T_{sa\mathbf{a}^{-i}s'}\frac{d\big[\sum_c X^i_{s'c}\mathcal{Q}^i_{s'c}(\mathbf{X})\big]}{dX^j_{rb}}. \tag{A18}$$

The derivative of $\sum_c X^i_{s'c}\mathcal{Q}^i_{s'c}(\mathbf{X})$ reads

$$\frac{d\big[\sum_c X^i_{s'c}\mathcal{Q}^i_{s'c}(\mathbf{X})\big]}{dX^j_{rb}} = \sum_c\left[\frac{dX^i_{s'c}}{dX^j_{rb}}\mathcal{Q}^i_{s'c}(\mathbf{X}) + X^i_{s'c}\frac{d\mathcal{Q}^i_{s'c}}{dX^j_{rb}}\right]. \tag{A19}$$

All remaining terms have already been given in the previous section for the Q learning Jacobian matrix.

3. Actor-critic learning

For the actor-critic learning update, Eq. (A3) reads

$$A^i_{sa} := X^i_{sa}\exp\big[\alpha^i\beta^i\hat{D}^i_{sa}(\mathbf{X})\big], \tag{A20}$$

with the truncated temporal difference error

$$\hat{D}^i_{sa}(\mathbf{X}) := (1-\gamma^i)_{\mathbf{TX}^{-i}}\langle R\rangle^i_{sa} + \gamma^{i\,\text{next}}\mathcal{V}^i_{sa}(\mathbf{X}). \tag{A21}$$

The derivative of the next value estimate is obtained by

$$\frac{d^{\text{next}}\mathcal{V}^i_{sa}(\mathbf{X})}{dX^j_{rb}} = \sum_{s'}\sum_{\mathbf{a}^{-i}}\frac{d\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}}{dX^j_{rb}}T_{sa\mathbf{a}^{-i}s'}V^i_{s'}(\mathbf{X}) + \sum_{s'}\sum_{\mathbf{a}^{-i}}\mathbf{X}^{-i}_{s\mathbf{a}^{-i}}T_{sa\mathbf{a}^{-i}s'}\frac{dV^i_{s'}(\mathbf{X})}{dX^j_{rb}}. \tag{A22}$$

The derivative of the next value $V^i_{s'}$ is given by Eq. (A13). These are all terms necessary to compute the Jacobian matrix for the actor-critic learning update.

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 1998).

[2] L. Busoniu, R. Babuska, and B. De Schutter, A comprehensive survey of multiagent reinforcement learning, IEEE Trans. Syst. Man Cybernet. C **38**, 156 (2008).

[3] M. Wiering and M. van Otterlo, *Reinforcement Learning: State-of-the-Art* (Springer Verlag, Berlin, 2012).

[4] A. Shah, Psychological and neuroscientific connections with reinforcement learning, in *Reinforcement Learning* (Springer, Berlin, 2012), pp. 507–537.

[5] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, Neuroscience-inspired artificial intelligence, Neuron **95**, 245 (2017).

[6] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*, Vol. 2 (MIT Press, Cambridge, MA, 1998).

[7] A. E. Roth and I. Erev, Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term, Games Econ. Behav. **8**, 164 (1995).

[8] I. Erev and A. E. Roth, Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria, Am. Econ. Rev. **88**, 848 (1998).

[9] C. Camerer and T. Hua Ho, Experienced-weighted attraction learning in normal form games, Econometrica **67**, 827 (1999).

[10] C. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton University Press, Princeton, NJ, 2003).

[11] W. B. Arthur, Designing economic agents that act like human agents: A behavioral approach to bounded rationality, Am. Econ. Rev. **81**, 353 (1991).

[12] W. B. Arthur, Complexity and the economy, Science **284**, 107 (1999).

[13] M. W. Macy and A. Flache, Learning dynamics in social dilemmas, Proc. Natl. Acad. Sci. USA **99**, 7229 (2002).

[14] J. G. Cross, A stochastic learning model of economic behavior, Quart. J. Econ. **87**, 239 (1973).

[15] T. Börgers and R. Sarin, Learning through reinforcement and replicator dynamics, J. Econ. Theory **77**, 1 (1997).

[16] M. Marsili, D. Challet, and R. Zecchina, Exact solution of a modified el farol's bar problem: Efficiency and the role of market impact, Physica A **280**, 522 (2000).

[17] Y. Sato, E. Akiyama, and J. D. Farmer, Chaos in learning a simple two-person game, in Ref. [18], pp. 4748–4751.

[18] Y. Sato and J. P. Crutchfield, Coupled replicator equations for the dynamics of learning in multiagent systems, in Ref. [15], p. 015206.

[19] Y. Sato, E. Akiyama, and J. P Crutchfield, Stability and diversity in collective adaptation, Physica D **210**, 21 (2005).

[20] T. Galla, Intrinsic Noise in Game Dynamical Learning, Phys. Rev. Lett. **103**, 198702 (2009).

[21] T. Galla, Cycles of cooperation and defection in imperfect learning, J. Stat. Mech.: Theory Exp. (2011) P08007.

[22] A. J. Bladon and T. Galla, Learning dynamics in public goods games, Phys. Rev. E **84**, 041132 (2011).

[23] J. Realpe-Gomez, B. Szczesny, L. Dall'Asta, and T. Galla, Fixation and escape times in stochastic game learning, J. Stat. Mech.: Theory Exp. (2012) P10022.

[24] J. B. T. Sanders, T. Galla, and J. L. Shapiro, Effects of noise on convergent game-learning dynamics, J. Phys. A: Math. Theor. **45**, 105001 (2012).

[25] T. Galla and J. D. Farmer, Complex dynamics in learning complicated games, Proc. Natl. Acad. Sci. USA **110**, 1232 (2013).

[26] A. Aloric, P. Sollich, P. McBurney, and T. Galla, Emergence of cooperative long-term market loyalty in double auction markets, PloS one **11**, e0154606 (2016).

[27] K. Tuyls, K. Verbeeck, and T. Lenaerts, A selection-mutation model for q-learning in multi-agent systems, in *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems* (ACM, New York, 2003), pp. 693–700.

[28] D. Bloembergen, K. Tuyls, D. Hennes, and M. Kaisers, Evolutionary dynamics of multi-agent learning: A survey, J. Artif. Intell. Res. **53**, 659 (2015).

[29] K. Tuyls and A. Nowé, Evolutionary game theory and multi-agent reinforcement learning, Knowl. Eng. Rev. **20**, 63 (2005).

[30] K. Tuyls, P. Jan'T Hoen, and B. Vanschoenwinkel, An evolutionary dynamical analysis of multi-agent learning in iterated games, Auton. Agents Multi-Agent Syst. **12**, 115 (2006).

[31] K. Tuyls and S. Parsons, What evolutionary game theory tells us about multiagent learning, Artif. Intell. **171**, 406 (2007).

[32] M. Kaisers and K. Tuyls, Frequency adjusted multi-agent q-learning, in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, Vol. 1 (International Foundation for Autonomous Agents and Multiagent Systems, 2010), pp. 309–316.

[33] D. Hennes, K. Tuyls, and M. Rauterberg, State-coupled replicator dynamics, in *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, Vol. 2 (International Foundation for Autonomous Agents and Multiagent Systems, 2009), pp. 789–796.

[34] P. Vrancx, K. Tuyls, and R. Westra, Switching dynamics of multi-agent learning, in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, Volume 1 (International Foundation for Autonomous Agents and Multiagent Systems, 2008), pp. 307–313.

[35] D. Hennes, M. Kaisers, and K. Tuyls, Resq-learning in stochastic games, in *Proceedings of the AAMAS Workshop on Adaptive and Learning Agents*, May 2010, Toronto, Canada (2010), p. 8.

[36] L. S. Shapley, Stochastic games, Proc. Natl. Acad. Sci. USA **39**, 1095 (1953).

[37] J.-F. Mertens and A. Neyman, Stochastic games, Int. J. Game Theory **10**, 53 (1981).

[38] E. Akiyama and K. Kaneko, Dynamical systems game theory and dynamics of games, Physica D **147**, 221 (2000).

[39] E. Akiyama and K. Kaneko, Dynamical systems game theory II: A new approach to the problem of the social dilemma, Physica D **167**, 36 (2002).

[40] M. T. J. Spaan, Partially observable markov decision processes, in *Reinforcement Learning* (Springer, Berlin, 2012), pp. 387–414.

[41] F. A. Oliehoek, Decentralized pomdps, in *Reinforcement Learning* (Springer, Berlin, 2012), pp. 471–503.

[42] R. Bellman, A Markovian decision process, Ind. Univ. Math. J. **6**, 679 (1957).

[43] S. Lange, T. Gabel, and M. Riedmiller, Batch reinforcement learning, in *Reinforcement Learning* (Springer, Berlin, 2012), pp. 45–73.

[44] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, Human-level control through deep reinforcement learning, Nature **518**, 529 (2015).

[45] C. Hilbe, Š. Šimsa, K. Chatterjee, and M. A. Nowak, Evolution of cooperation in stochastic games, Nature **559**, 246 (2018).

[46] M. Sandri, Numerical calculation of Lyapunov exponents, Math. J. **6**, 78 (1996).

[47] A. Bandura, *Social Learning Theory* (Prentice-Hall, Upper Saddle River, NJ, 1977), pp. 15–55.

[48] M. Smolla, R. T. Gilman, T. Galla, and S. Shultz, Competition for resources can explain patterns of social and individual learning in nature, Proc. Roy. Soc. Lond. B **282** 20151405 (2015).

[49] W. Barfuss, J. F. Donges, M. Wiedermann, and W. Lucht, Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution, Earth Syst. Dynam. **8**, 255 (2017).

[50] S. Banisch and E. Olbrich, Opinion polarization by learning from social feedback, J. Math. Soc. **43**, 76 (2019).

[51] S. A. Levin, The mathematics of sustainability, Not. Am. Math. Soc. **60**, 392 (2013).

[52] J. F. Donges, R. Winkelmann, W. Lucht, S. E. Cornell, J. G. Dyke, J. Rockström, J. Heitzig, and H. J. Schellnhuber, Closing the loop: Reconnecting human dynamics to Earth System science, Anthrop. Rev. **4**, 151 (2017).

[53] R. M. Dawes, Social dilemmas, Annu. Rev. Psychol. **31**, 169 (1980).

[54] J. Heitzig, T. Kittel, J. F. Donges, and N. Molkenthin, Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dynam. **7**, 21 (2016).

[55] E. Lindkvist and J. Norberg, Modeling experiential learning: The challenges posed by threshold dynamics for sustainable renewable resource management, Ecol. Econ. **104**, 107 (2014).

[56] C. Schill, T. Lindahl, and A.-S. Crépin, Collective action and the risk of ecosystem regime shifts: Insights from a laboratory experiment, Ecol. Soc. **20**, 48 (2015).

[57] M. Milinski, R. D. Sommerfeld, H.-J. Krambeck, F. A. Reed, and J. Marotzke, The collective-risk social dilemma and the

prevention of simulated dangerous climate change, Proc. Natl. Acad. Sci. USA **105**, 2291 (2008).

[58] S. Barrett and A. Dannenberg, Climate negotiations under scientific uncertainty, Proc. Natl. Acad. Sci. USA **109**, 17372 (2012).

[59] W. Barfuss, J. F. Donges, S. J. Lade, and J. Kurths, When optimization for governing human-environment tipping elements is neither sustainable nor safe, Nat. Commun. **9**, 2354 (2018).

[60] J. F. Donges and W. Barfuss, From math to metaphors and back again: Social-ecological resilience from a multi-agent-environment perspective, GAIA **26**, 182 (2017).

[61] W. Barfuss, wbarfuss/DetRL: Release for reference in publication, Zenodo (2018), http://doi.org/10.5281/zenodo.1495091.

PHYSICAL REVIEW E **101**, 042311 (2020)

Dynamics of tipping cascades on complex networks

Jonathan Krönke,[1,2,*] Nico Wunderling,[1,2,3,†] Ricarda Winkelmann,[1,2] Arie Staal,[4] Benedikt Stumpf,[1,5] Obbe A. Tuinenburg,[4,6] and Jonathan F. Donges[1,4,‡]

[1]*Earth System Analysis, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, 14473 Potsdam, Germany*
[2]*Institute of Physics and Astronomy, University of Potsdam, 14476 Potsdam, Germany*
[3]*Department of Physics, Humboldt University of Berlin, 12489 Berlin, Germany*
[4]*Stockholm Resilience Centre, Stockholm University, 10691 Stockholm, Sweden*
[5]*Department of Physics, Free University Berlin, 14195 Berlin, Germany*
[6]*Copernicus Institute, Faculty of Geosciences, Utrecht University, 3584 CB Utrecht, The Netherlands*

(Received 9 May 2019; revised manuscript received 15 January 2020; accepted 18 March 2020; published 29 April 2020)

Tipping points occur in diverse systems in various disciplines such as ecology, climate science, economy, and engineering. Tipping points are critical thresholds in system parameters or state variables at which a tiny perturbation can lead to a qualitative change of the system. Many systems with tipping points can be modeled as networks of coupled multistable subsystems, e.g., coupled patches of vegetation, connected lakes, interacting climate tipping elements, and multiscale infrastructure systems. In such networks, tipping events in one subsystem are able to induce tipping cascades via domino effects. Here, we investigate the effects of network topology on the occurrence of such cascades. Numerical cascade simulations with a conceptual dynamical model for tipping points are conducted on Erdős-Rényi, Watts-Strogatz, and Barabási-Albert networks. Additionally, we generate more realistic networks using data from moisture-recycling simulations of the Amazon rainforest and compare the results to those obtained for the model networks. We furthermore use a directed configuration model and a stochastic block model which preserve certain topological properties of the Amazon network to understand which of these properties are responsible for its increased vulnerability. We find that clustering and spatial organization increase the vulnerability of networks and can lead to tipping of the whole network. These results could be useful to evaluate which systems are vulnerable or robust due to their network topology and might help us to design or manage systems accordingly.

DOI: 10.1103/PhysRevE.101.042311

I. INTRODUCTION

In the last decades the study of tipping elements has become a major topic of interest in climate science. Tipping elements are subsystems of the Earth system that may pass a critical threshold (tipping point) at which a tiny perturbation can qualitatively alter the state or development of the subsystem [1]. However, tipping points also occur in various complex systems such as systemic market crashes in financial markets [2], technological innovations [3], or shallow lakes [4] and other ecosystems [5]. Understanding their dynamics is thus crucial not only for climate science but also for other disciplines that use complex systems approaches.

Many tipping elements are not independent of each other [6]. In such cases, if one tipping element passes its tipping point, the probability of tipping of a second tipping element is often increased [7], yielding the potential of tipping cascades [8] via domino effects with significant potential impacts on human societies in the case of climate tipping elements [9]. In this study, we investigate the dynamics of complex networks of interacting tipping elements. A tipping element is described by a differential equation based on the normal form of the cusp catastrophe, which exhibits fold bifurcations and hysteresis properties. The interactions are accounted for by linear coupling terms. Many environmental tipping points can be described as fold bifurcations [10] and prototypical conceptual models that exhibit fold bifurcations have been developed for the thermohaline circulation [11], the Greenland ice sheet [12], and tropical rainforests [13] among others. Coupled cusp catastrophes have been studied in detail for two or three subsystems [6,14,15] or in combination with Hopf bifurcations [16]. On the other hand, threshold models for global cascades on large random networks have been investigated [17].

Here, we study cascades in complex systems with continuous state space that are moderate in size yet large enough for statistical properties of the complex interaction networks to become relevant. Cascades in complex systems with continuous state space have been investigated, for example, for power grids [18,19]. We use a paradigmatic coupled hysteresis model based on the normal form of the cusp catastrophe. Employing different network topologies such as Erdős-Rényi (ER), Watts-Strogatz (WS), and Barabási-Albert (BA) networks as

[*] kroenke@pik-potsdam.de

[†] Author to whom correspondence should be addressed: wunderling@pik-potsdam.de

[‡] Author to whom correspondence should be addressed: donges@pik-potsdam.de

well as networks generated from moisture-flow data on the Amazon rainforest, we investigate the effect of topological properties of the network. We find that networks with a large average clustering coefficient are more vulnerable to cascading tipping and discuss how this is connected to the occurrence of small-scale motifs such as direct feedback and feed-forward loops. We consistently observe that networks with spatial organization like the small-world and Amazon networks are more vulnerable than strongly disordered networks.

II. THE MODEL

A. System

In our conceptual model, a tipping element is represented by a (real) time-dependent quantity $x(t)$ that evolves according to the autonomous ordinary differential equation

$$\frac{dx}{dt} = -a(x-x_0)^3 + b(x-x_0) + r, \tag{1}$$

where r is the control parameter and $a, b > 0$. The parameters a and b control the strength of these effects, respectively, and x_0 controls the position of the system on the x axis. The equation thus has one stable equilibrium for $|r| > r_{\text{crit}}$ and a bistable region for $-r_{\text{crit}} < r < r_{\text{crit}}$ (see the bifurcation diagram depicted in the box in Fig. 1).

We describe the characteristic behavior of Eq. (1): If the system state is initially in the lower stable equilibrium ($x \approx 0$) and r is slowly increased, eventually at $r = r_{\text{crit}}$ a tipping point is reached and a critical transition to the upper stable equilibrium ($x \approx 1$) occurs. If r is afterwards decreased, the system state stays on the upper branch and, only at $r = -r_{\text{crit}}$, tips down to the lower branch again. Equation (1) is a minimal model for ecosystems with alternative stable states and hysteresis [5] but can also be used to conceptualize other systems with similar properties such as the thermohaline circulation and ice sheets [12,20].

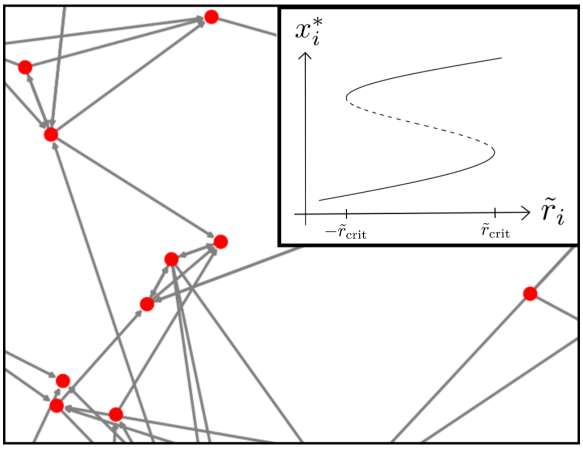


FIG. 1. Illustration of a tipping network. Each node represents a tipping element with a corresponding state variable x_i. A directed link corresponds to a positive linear coupling with strength d. The effective control parameter $\tilde{r}_i$ of a node depends on the state of the nodes it is coupled to. The equilibria with respect to the effective control parameter are qualitatively illustrated in the box.

Next, we consider a directed network of N interacting tipping elements as a linearly coupled system of ordinary differential equations,

$$\frac{dx_i}{dt} = -a(x_i - x_0)^3 + b(x_i - x_0) + \underbrace{r_i + d\sum_{j=1,j\neq i}^{N} a_{ij}x_j}_{\tilde{r}_i(x_1,x_2,\ldots,x_N)}, \tag{2}$$

where $d > 0$ is the coupling strength and

$$a_{ij} = \begin{cases} 1 & \text{if there is a directed link from element } j \text{ to element } i, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

For simplicity, we use the same parameters a and b for all tipping elements in the network. An illustration of such a system with several tipping elements is depicted in Fig. 1. Similar systems have been studied with diffusive coupling focusing on hysteresis effects [21].

We briefly review the behavior of two tipping elements with unidirectional coupling ($X_1 \rightarrow X_2$) [6]. The elements of the adjacency matrix are $a_{21} = 1$ and $a_{12} = 0$, which means that element 1 has an effect on element 2 but there is no effect in the other direction. As r_1 is slowly increased, it approaches its tipping point at r_{crit} and eventually tips from x_- to x_+. The effective control parameter $\tilde{r}_2$ is thus increased by $\Delta\tilde{r} = d(x_+ - x_-)$. For $r_2 = 0$, a tipping event in the second element is induced if $\Delta\tilde{r} > r_{\text{crit}}$ and therefore if the coupling strength exceeds a critical threshold of $d_{\text{c}} = \frac{r_{\text{crit}}}{x_+ - x_-}$.

B. Network models

To investigate the effect of the network topology on tipping cascades we use different network models: We use three well-known models, the Erdős-Rényi model [22], the Watts-Strogatz model [23], and the Barabási-Albert model [24]. We slightly extend the latter two models such that we are able to generate and compare directed networks with a controllable average degree $\langle k \rangle = \langle k_{\text{in}} + k_{\text{out}} \rangle$. Furthermore, we use models to control the reciprocity and average clustering coefficient as well as a directed configuration model and a stochastic block model. All network models are briefly discussed in the following paragraphs.

(i) The ER model is a simple random network model, where a directed link between two elements i and j is added with probability p. The resulting average degree is $\langle k \rangle \approx p(N-1)$.

Reprinted paper with permission from J. Krönke et al., Dynamics of tipping cascades on complex networks, Phys. Rev. E, vol. 101, no. 4, pp. 1-19, 2020, doi: 10.1103/PhysRevE.101.042311. Copyright 2020 by the American Physical Society

(ii) The WS model is usually used to generate networks with large clustering coefficients but small average path lengths to resemble the small-world phenomenon [25]. We implement a directed WS model as follows: Initially, a regular network is generated where each node i is connected in both directions to its m nearest neighbors, e.g., nodes $i+1,\ i-1,\dots,i+\frac{m}{2},\ i-\frac{m}{2}$. Therefore, m has to be an even integer and the average degree of the resulting regular network is equal to m. In order to generate networks with arbitrary average degree, m is chosen such that the average degree of the resulting regular network is larger than the desired average degree. Then, until the average degree of the network matches the desired average degree, links are randomly deleted. Finally, each of the remaining links is rewired with probability β, similar to the usual WS model [23]. With increasing rewiring probability β the generated network becomes more and more random.

(iii) The BA model is used to generate scale-free networks, i.e., networks with a power-law degree distribution. We implement a directed BA model as follows: We start with two bidirectionally coupled nodes. Every additional node is in both directions connected to an already existing node i with probability $p=\frac{k_i^{\text{in}}+k_i^{\text{out}}}{\sum_{m,n}a_{mn}}$. When the specified network size N is reached, the average degree $\langle k\rangle\approx\frac{\sum_{m,n}a_{mn}}{N}$ is compared to the desired average degree. If the average degree is smaller than the desired average degree, links between randomly selected nodes i and j are added with probability $p=\frac{k_i^{\text{in}}+k_i^{\text{out}}+k_j^{\text{in}}+k_j^{\text{out}}}{2\sum_{m,n}a_{mn}}$ until the average degree matches the desired average degree. Otherwise, if the average degree is greater than the desired average degree, links are randomly deleted as in the WS model.

(iv) To generate networks with arbitrary reciprocity R, we initially generate an ER network where all links are reciprocal ($R=1$). Afterwards, links are randomly chosen and rewired until the desired reciprocity is achieved.

(v) The procedure to generate networks with arbitrary average clustering coefficent $\mathcal{C}$ is similar. Initially a network with only reciprocal triangles between three randomly chosen nodes is generated. Afterwards links are randomly chosen and rewired again until the desired average clustering coefficient is achieved. That way, we are able to generate networks with an average clustering coefficient between $\mathcal{C}=0.05$ and $\mathcal{C}=0.35$. Note that the reciprocity is also large for networks with a large average clustering coefficient.

(vi) A directed configuration model can be used to generate networks with arbitrary average in and out degree. Links are randomly assigned to node pairs where the corresponding in and out degree has not been reached before [26].

(vii) Finally, stochastic block models (SBMs) are used to generate networks with community structures. For each (directed) combination of communities there is a separate link probability, which is usually high within the community and low between two communities [27].

C. Simulation procedure

We use the system given in Eq. (2) and conduct cascade simulations on different network topologies. The parameters of the equation are chosen such that $r_{\text{crit}}=0.183$ and for $r=0$ the two stable equilibria are $x_-=0$ and $x_+=1$ for all elements. The resulting parameters are $a=4$, $b=1$, and $x_0=0.5$. Consider a network with N tipping elements and a topology that is described by the adjacency matrix $A=(a_{ij})$. Initially, $r_i=0$ and $x_i=0$ for all $i=1,\dots,N$. The algorithm of a cascade simulation is the following:

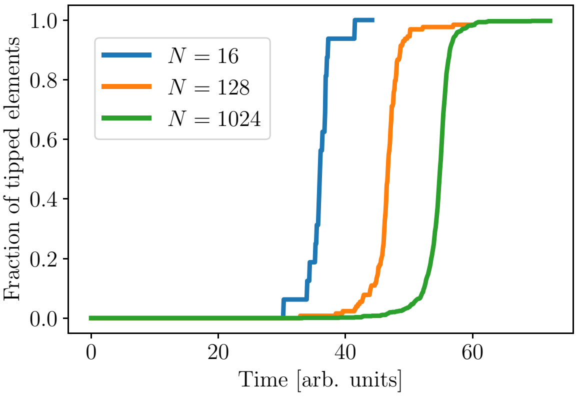


FIG. 2. Cascade simulations on ER networks of different sizes, an average degree of $\langle k\rangle\approx 5$, and a coupling strength of $d=0.2$. The time evolution of the fraction of tipped elements is shown.

(1) Choose a random starting node m of the network.

(2) Slowly increase r_m ($r_m\rightarrow r_m+\Delta r$).

(3) Let the system equilibrate, e.g., integrate the ODE system until $\dot{x}_i<\varepsilon$ for all $i=1,\dots,N$.

(4) Check whether at least one element tipped. If not, jump back to step 2. Otherwise, count the total number of tipped elements.

The algorithm stops when the starting node m tips, which is always the case. We normalize the total number of tipped elements (minus 1 for the starting node) by the number of nodes that can be reached on a directed path from the starting node (the size of the out component). We call the resulting number cascade size L. Note that due to the normalization a small disconnected component where all elements tip is also considered as a cascade with size $L=1$ even though only a small number of elements was tipped. The ODE system was integrated with the function `scipy.integrate.odeint` from the SCIPY python package [28]. In all simulations, $\Delta r=0.01$ and $\varepsilon=0.005$ were used. Examples of tipping cascades with size $L=1$ are shown in Fig. 2 for ER networks with different-sized N.

III. RESULTS AND DISCUSSION

A. Cascades on generic network topologies

We start with cascade simulations on networks generated with the ER model. For any parameter combination we generate 100 different networks and simulate one cascade on each network. We use the average cascade size from these simulations as a measure of the vulnerability of the corresponding network structure, ranging from robust ($\langle L\rangle=0$) to highly vulnerable ($\langle L\rangle=1$) networks. The dependence of the average cascade size with respect to the coupling strength is shown in the upper panel in Fig. 3 for random networks with

Reprinted paper with permission from J. Krönke et al., Dynamics of tipping cascades on complex networks, Phys. Rev. E, vol. 101, no. 4, pp. 1-19, 2020, doi: 10.1103/PhysRevE.101.042311. Copyright 2020 by the American Physical Society

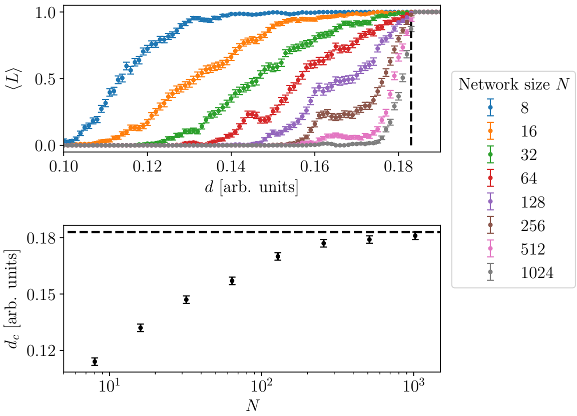


FIG. 3. Network size dependency of critical coupling strength in ER networks with $\langle k\rangle \approx 5$. Upper panel: Average cascade size with respect to the coupling strength in the transition region. Each average is calculated from 100 cascade simulations on different randomly generated networks with $N = 100$. Error bars indicate the standard error. Lower panel: Approximate critical coupling strength (coupling strength where $\langle L\rangle \approx 0.5$) with respect to the network size N. The dashed line indicates the critical coupling strength $d_c \approx r_{\text{crit}} = 0.183$ for a simple unidirectional coupling of two elements.

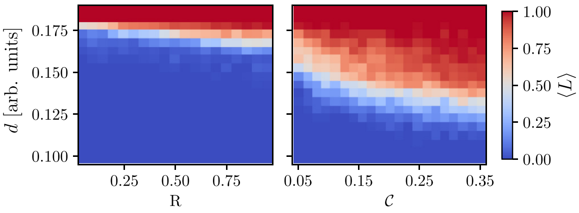


FIG. 4. Dependence of the transition region on the reciprocity R (left panel) and on the clustering coefficient $\mathcal{C}$ (right panel). Each average is calculated from 100 cascade simulations on different randomly generated networks with $N = 100$.

a fixed average degree $\langle k\rangle \approx 5$. For low coupling strengths ($d \lesssim 0.1$) the network is not affected by the externally induced tipping of one element and the average cascade size remains 0. With increasing coupling strength, a transition from robust to vulnerable networks is observed. From the analysis of the unidirectional system, a sharp transition at $d \approx r_{\text{crit}}$ would be expected for all networks. However, only for $N \to \infty$ does the transition become more and more steep and approximately approach r_{crit}. For networks of finite size, the onset of the transition is shifted to lower coupling strengths with decreasing network size. We hypothesize that the reason for this is two effects: The first effect is the destabilization of the system by feedback loops ($X_1 \leftrightarrows X_2$), which can lead to a decrease in the tipping point r_{crit} of certain nodes. The second effect is due to the gradual change in the state of a tipping element X_3 that is coupled to another element ($X_1 \to X_3$). When the element X_1 tips, the state of the element X_3 will be slightly altered even if it does not tip. If it is coupled to another element X_2, however ($X_2 \to X_3$), the effective control parameter of element X_3 will be slightly increased, by an increment of the order $\Delta\tilde{r} \sim d^2$. Therefore an additional indirect coupling with one intermediate node, called a feed-forward loop, will decrease the critical coupling strength d_c of the target node. But how can the size dependence of the critical coupling strength be explained? The reason for this is the following: With increasing network size while fixing the average degree, the relative density of the motifs decreases, and thus, for $N \to \infty$, the destabilizing effect of the motifs vanishes. Therefore, the critical coupling strength d_c approaches the critical coupling strength of a unidirectionally coupled system. If, in contrast, we fixed the link density, the relative density of motifs would increase and thus the critical coupling strength would probably decrease with increasing network size.

To test this hypothesis, cascade simulations on networks with different reciprocities and average clustering coefficients are conducted. The reciprocity is the number of reciprocated links ($a_{ij} = a_{ji} = 1$) divided by the total number of links in the network. Thus, the reciprocity measures the relative amount of feedback loops in the tipping network. The average clustering coefficient is the number of triangles a node is part of divided by the potential number of triangles averaged over all nodes [29]. Therefore, the average clustering coefficient is strongly related to the number of feed-forward loops. Simulation results for different reciprocities R can be seen in the left panel in Fig. 4. As expected, for networks with a high reciprocity, the transition region is shifted to lower coupling strengths. As can be seen, however, the dependence on the reciprocity is rather weak. Simulation results for networks with different average clustering coefficient $\mathcal{C}$ are shown in the right panel in Fig. 4. It can be clearly seen that the vulnerability to tipping cascades is significantly increased for high average clustering coefficients. There are eight motifs that contribute to the average clustering coefficient in a directed network, two (indirect) feedback loops and six feed-forward loops [30]. We suspect that the effect of indirect feedback loops is smaller than the effect of direct feedback loops for $d < 1$. Therefore, we conclude that feed-forward loops are mainly responsible for the increased vulnerability of networks with large average clustering (see Fig. 4).

We also observe a transition of the average cascade size when the coupling strength is held constant at $d = 0.15$ and the average degree is varied (Fig. 5). In this case the transition is shifted to higher average degrees when the network size increases, because a higher average degree is necessary to yield the same relative density of destabilizing motifs.

Cascade distributions for $\langle k\rangle \approx 5$ and selected coupling strengths at the onset, in the center, and at the end of the respective transition region are shown in Fig. 6. We find a bimodal distribution of very small cascades ($L \approx 0$) and very large cascades ($L \approx 1$). For networks with small-world and scale-free topology generated with the WS model with $\beta = 0.1$ and the BA model, respectively, we observe similar transitions of the average cascade size. For the scale-free topology, the large cascades are distributed around an average size $\langle L\rangle < 1$. This can be explained by the preferential attachment mechanism. Through this mechanism a large number of weakly connected elements develop which can only be tipped when the coupling strength is very high ($d \gtrsim r_{\text{crit}}$).

Reprinted paper with permission from J. Krönke et al., Dynamics of tipping cascades on complex networks, Phys. Rev. E, vol. 101, no. 4, pp. 1-19, 2020, doi: 10.1103/PhysRevE.101.042311. Copyright 2020 by the American Physical Society

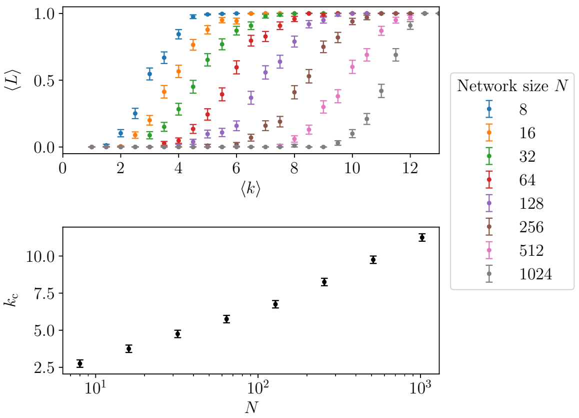


FIG. 5. Network size dependency of the critical average degree k_c in ER networks with $d = 0.15$. Upper panel: Average cascade size with respect to the average degree in the transition region. Each average is calculated from 100 cascade simulations on different randomly generated networks with $N = 100$. Error bars indicate the standard error in both panels. Lower panel: Approximate critical average degree (average degree where $\langle L \rangle \approx 0.5$) with respect to the network size N.

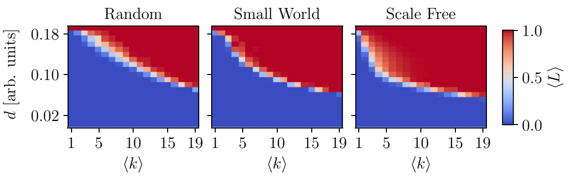


FIG. 7. Average cascade size $\langle L \rangle$ with respect to average degree $\langle k \rangle$ and coupling strength d for three network topologies. Random networks generated with the ER model (left), small-world topology networks generated with the WS model and $\beta = 0.1$ (center), and scale-free networks generated with the BA model (right). Each average is calculated from 100 cascade simulations on different randomly generated networks with $N = 100$.

Now we focus on the effect of the network topology. For all network models, the transition from robust to vulnerable networks is shifted to lower coupling strengths when the average degree is increased (Fig. 7). The topology of the network has a significant effect on this shift of the transition region for sparse networks ($\langle k \rangle \approx 5$). For networks with small-world and scale-free topology, the transition is shifted to lower coupling strengths compared to the simple random topology generated with the ER model. For the scale-free topology the transition width is also significantly increased for $\langle k \rangle \approx 5$. For denser networks ($\langle k \rangle \gtrsim 19$), the differences between the network topologies are less pronounced.

We further investigate in which way the rewiring in the WS model decreases the vulnerability of the network. In Fig. 8 the shift of the transition region to higher coupling strengths with respect to the rewiring probability β can be clearly seen. The increase in the critical coupling strength mainly occurs between $\beta = 0.1$ and $\beta = 1$. The lower panel in the figure again demonstrates how this corresponds to the decay of the average clustering coefficient $\mathcal{C}$. Thus, we again conclude that tipping networks with an increased average clustering coefficient such as small-world networks (but also spatially

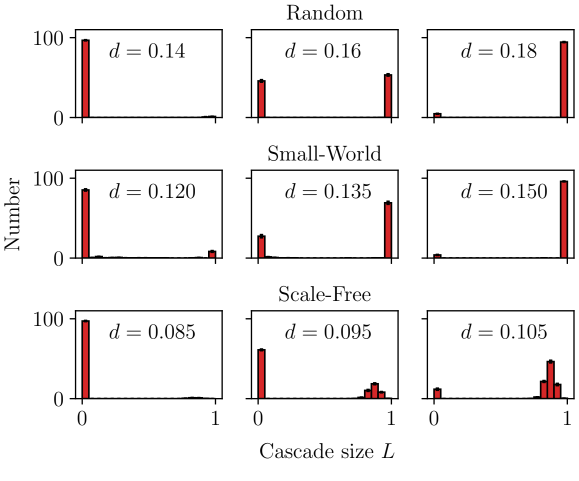


FIG. 6. Distributions of cascade sizes L for different network topologies. A random topology generated with the ER model (first row), a small-world topology generated with the WS model and $\beta = 0.1$ (second row), and a scale-free topology generated with the BA model (third row). Each distribution is an average of 10 distributions with 100 cascade simulations on different networks with $N = 100$ and $\langle k \rangle \approx 5$. The (almost-invisible) error bars indicate the standard error across the 10 distributions. Three coupling strengths for each network topology are shown: one where almost no cascades occur; one where in about half of the simulations cascades are triggered; and one where in almost all simulations cascades are triggered.

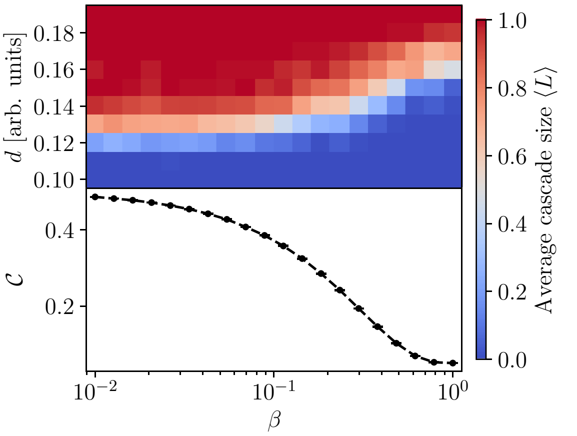


FIG. 8. Shift of the transition (upper panel) and average clustering coefficient $\mathcal{C}$ (lower panel) with increasing rewiring probability β for WS networks with $N = 100$ and $\langle k \rangle \approx 5$. The shift of the transition towards higher coupling strengths for high rewiring probabilities corresponds to the decrease in the average clustering coefficient. The extent of the small black circles in the lower panel exceeds the standard error, which is therefore not visible.

Reprinted paper with permission from J. Krönke et al., Dynamics of tipping cascades on complex networks, Phys. Rev. E, vol. 101, no. 4, pp. 1-19, 2020, doi: 10.1103/PhysRevE.101.042311. Copyright 2020 by the American Physical Society

structured networks [31]; see Sec. III B) are especially vulnerable to cascades and that the average clustering coefficient is a good indicator of the vulnerability of a network topology.

B. Cascades on spatial network topologies from moisture-flow data

To investigate the effects of spatial organization of the network on vulnerability with respect to tipping cascades, we apply our model to network topologies generated from data of atmospheric moisture flows between different forest cells in the Amazon. On a local scale, the Amazon may exhibit alternative stable states between rainforest and savanna, with tipping points between them depending on rainfall levels [32–35]. Models that capture the basic mechanisms also reveal a bifurcation structure with hysteresis and saddle-node bifurcations with rainfall level as the control parameter, comparable to our conceptual model [36]. On a regional scale, the forest enhances rainfall through the "transpiration" of groundwater to the atmosphere; local-scale tipping may thus increase the vulnerability of remote forest patches by allowing less local precipitation to be passed on to other patches because the transpiration capacity of savanna is lower than that of forest. Therefore, the Amazon can be thought of as a spatial network of local-scale tipping elements. Note that the Amazon as a whole is often viewed as a tipping element [37]. In our framework, vulnerable regimes where tipping of single cells induces large cascades correspond to such threshold behavior of the large-scale Amazon system. Complex-network approaches such as a cascade model inspired by the Watts model [17] have been applied to observation-based data on Amazon forest patches [38]. Here we analyze the effect of the network structure of transpired-moisture flows for the Amazon that were calculated by Staal *et al.* [39], aggregated to a single year (2014) on 1° spatial resolution.

As our analysis is focused on the effect of the network topology, we neglect the actual moisture-flow values and use a homogeneous coupling strength analogous to the above simulations. This makes the simulation results less realistic and applicable, however, we do not aim to draw conclusions about the Amazon system. Rather, we want to compare the network topology to common random networks and identify topological effects on the vulnerability of tipping networks with respect to tipping cascades.

To generate and compare networks with arbitrary average degree, similar to the random network topologies above, we calculate a moisture-flow threshold from a specified average degree. Only when the moisture flow between two cells exceeds the threshold are these cells connected with a link in the corresponding direction. If a large average degree is specified, the threshold becomes small and the resulting network will be dense. That way we are able to generate networks with an arbitrary average degree from the data. An example network with $\langle k \rangle = 5$ is depicted in Fig. 9.

The average cascade size is calculated by conducting one cascade simulation with each node of the generated network as the starting node and averaging over the cascade size. We generate networks from data with a $1 \times 1^\circ$ grid ($N = 567$) and with a $2 \times 2^\circ$ grid ($N = 160$) and $\langle k \rangle = 5$. The average cascade size of ER networks with the same size is shown for comparison (upper panel in Fig. 10). For the Amazon network, the onset of the transition from robust to vulnerable networks is shifted to the lower coupling strength of $d \approx 0.08$ compared to the ER network. In contrast to the ER networks there is no strong size dependency. However, a small shift to lower coupling strengths is observed.

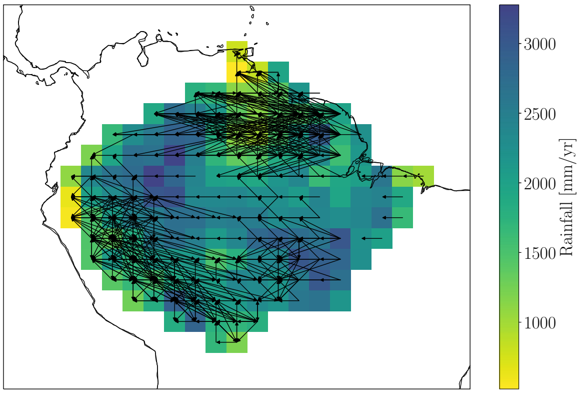


FIG. 9. Spatially organized network generated from atmospheric moisture-flow data ($2 \times 2^\circ$-grid resolution) of the Amazon rainforest. The threshold is chosen such that $\langle k \rangle = 5$. Total rainfall values for each node in 2014 are shown in the background.

Additionally to the Amazon moisture-flow network obtained by thresholding, we generate networks with a directed configuration model [26] and a stochastic block model [27] to isolate the effects of the degree sequence and the community

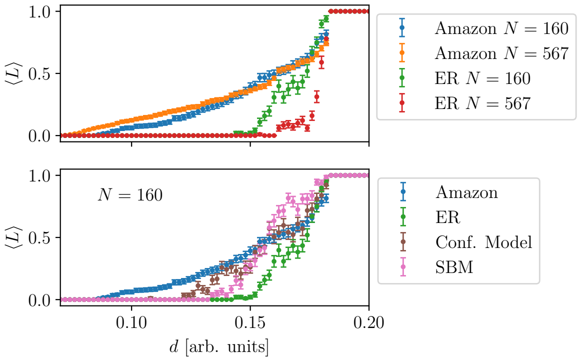


FIG. 10. Average cascade size $\langle L \rangle$ with respect to coupling strength for different networks with an average degree of $\langle k \rangle = 5$. Upper panel: Results for the networks generated from the moisture-flow data with $1 \times 1^\circ$-grid resolution (567 nodes) and $2 \times 2^\circ$-grid resolution (160 nodes). For comparison, simulation results for ER networks with the same network sizes are shown. Lower panel: Simulation results for a directed configuration model and a stochastic block model are compared with the results of the Amazon network and the ER networks with $N = 160$ for all networks. Error bars indicate the standard error. Note that the standard errors for the original moisture-flow networks are smaller than for the other network types. The reason is that all moisture-flow simulation results are based on the same network, whereas the other results are based on different randomly generated networks.

Reprinted paper with permission from J. Krönke et al., Dynamics of tipping cascades on complex networks, Phys. Rev. E, vol. 101, no. 4, pp. 1-19, 2020, doi: 10.1103/PhysRevE.101.042311. Copyright 2020 by the American Physical Society

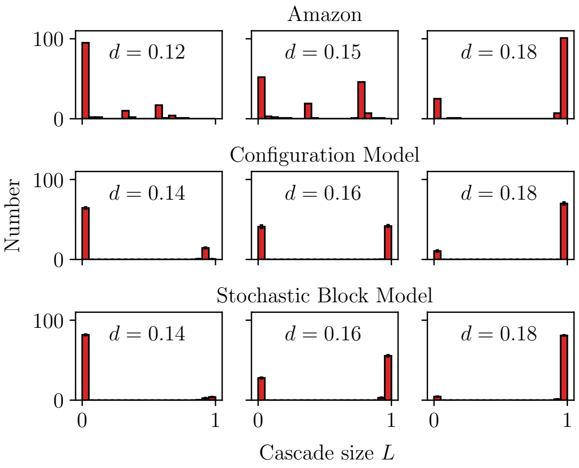


FIG. 11. Distribution of cascade sizes analogous to the above distributions for different networks generated from moisture-flow simulations of the Amazon rainforest ($N = 160$). Note that there is no standard error indicated (error bars) for the original moisture-flow networks, as there is only one distribution due to the deterministic network generation procedure.

structure of the network, respectively. For the directed configuration model, we specify the joint degree sequence of the Amazon network. For the SBM, we apply a Girvan-Newman algorithm to the original Amazon network [40]. The algorithm progressively removes edges with the highest edge betweenness, i.e., those rare links that connect separate communities. When the network breaks into two components, we calculate the elements of the probability matrix (fraction of links over possible links for the corresponding combination of components). With the probability matrix and the component sizes, we then generate a random network with the SBM.

In the lower panel in Fig. 10, the transition of the configuration model and the SBM is compared to the original Amazon network and the ER network with $N = 160$. Although the vulnerability of the network is increased in both cases compared to the ER model, neither of the topological properties alone, degree sequence or community structure, sufficiently explains the early onset of the transition in the original Amazon network.

Cascade distributions for the coarse resolution ($2 \times 2^\circ$ grid) are depicted in Fig. 11. They show that already for values of $d \approx 0.1$, cascades with two typical cascade sizes occur for the original Amazon network. With increasing coupling strength the frequency of these cascades increases and the cascade size is shifted to higher values. Comparing this observation to the network in Fig. 9 suggests that these cascades correspond to the two subclusters in the north and southwest regions of the Amazon rainforest. These subregions form clusters that are much more strongly connected than the rest of the network and are thus much more vulnerable to tipping cascades. Interestingly, separate tipping of subclusters is not observed for the networks generated with the SBM, implying that some relevant topological property of the spatially structured Amazon network, for example, the anisotropy of the link direction due to atmospheric wind patterns, might still be missing. The robust and vulnerable regimes of the networks are shown in Fig. 12. Consistent with the above results, we observe a shift of the transition to lower coupling strengths with increasing average degree $\langle k \rangle$ where the transition is smooth for the Amazon network and steep for the configuration model and the SBM. Similarly to the random network topologies, the differences are only relevant for the sparse regime below $\langle k \rangle \lesssim 19$.

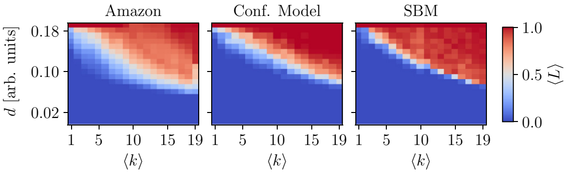


FIG. 12. Average cascade size $\langle L \rangle$ with respect to average degree and coupling strength for different networks generated with moisture-flow simulations of the Amazon rainforest ($N = 160$).

IV. CONCLUSION

The aim of our study was to assess the effect of the network topology on the vulnerability of tipping networks to cascades. This is not only important for understanding the effect that the tipping of potential tipping elements in the climate system might have on the complete Earth system, but also of high relevance for other fields that use complex system approaches. We found that networks with large average clustering coefficients and spatially structured networks are more vulnerable to tipping cascades than more disordered network topologies. This implies that the risk of a cascade's being triggered could be surprisingly high for real-world networks where large clustering is common. Furthermore, we found that the effect of the network topology is relevant only for relatively sparsely connected networks. The analysis of the Amazon network suggests that the structure of the forest-climate system in the Amazon might yield subregions that are especially vulnerable to tipping cascades. A detailed study using actual moisture flows could investigate the question whether the Amazon rainforest consists of separate subregional-scale tipping elements. Generally, heterogeneity in the parameters, for example, the temporal and spatial scales or the coupling strengths of the ODE system stated in Eq. (2), could have a further influence on the results [41].

ACKNOWLEDGMENTS

The authors thank M. Wiedermann and J. Heitzig for fruitful discussions. N.W. acknowledges support from IRTG 1740/TRP 2015/50122-0 funded by the DFG and FAPESP. N.W. is grateful for support through a scholarship from the Studienstiftung des deutschen Volkes. R.W. and J.F.D. are thankful for support from the Leibniz Association (project DominoES). A.S. and J.F.D. acknowledge support from the European Research Council project Earth Resilience in the Anthropocene (743080 ERA). A.S. and O.A.T. thank the Bolin Centre for Climate Research for support. O.A.T. acknowledges funding from the Netherlands Organization for

Reprinted paper with permission from J. Krönke et al., Dynamics of tipping cascades on complex networks, Phys. Rev. E, vol. 101, no. 4, pp. 1-19, 2020, doi: 10.1103/PhysRevE.101.042311. Copyright 2020 by the American Physical Society

Scientific Research Innovational Research Incentive Schemes VENI (No. 016.171.019). J.F.D. thanks the Stordalen Foundation (via the Planetary Boundaries Research Network PB.net) and the Earth League's EarthDoc program for financial support. The authors gratefully acknowledge the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research, and the Land Brandenburg for supporting this project by providing resources on the high-performance computer system at the Potsdam Institute for Climate Impact Research.

[1] T. M. Lenton, H. Held, E. Kriegler, J. W. Hall, W. Lucht, S. Rahmstorf, and H. J. Schellnhuber, Tipping elements in the earth's climate system, Proc. Natl. Acad. Sci. USA **105**, 1786 (2008).

[2] R. M. May, S. A. Levin, and G. Sugihara, Ecology for bankers, Nature **451**, 893 (2008).

[3] P. A. Herbig, A cusp catastrophe model of the adoption of an industrial innovation, J. Prod. Innov. Manage. **8**, 127 (1991).

[4] M. Scheffer and E. H. van Nes, Shallow lakes theory revisited: Various alternative regimes driven by climate, nutrients, depth and lake size, Hydrobiologia **584**, 455 (2007).

[5] M. Scheffer, S. Carpenter, J. A. Foley, C. Folke, and B. Walker, Catastrophic shifts in ecosystems, Nature **413**, 591 (2001).

[6] C. D. Brummitt, G. Barnett, and R. M. D'Souza, Coupled catastrophes: Sudden shifts cascade and hop among interdependent systems, J. R. Soc., Interface **12**, 20150712 (2015).

[7] E. Kriegler, J. W. Hall, H. Held, R. Dawson, and H. J. Schellnhuber, Imprecise probability assessment of tipping points in the climate system, Proc. Natl. Acad. Sci. USA **106**, 5041 (2009).

[8] W. Steffen, J. Rockström, K. Richardson, T. M. Lenton, C. Folke, D. Liverman, C. P. Summerhayes, A. D. Barnosky, S. E. Cornell, M. Crucifix, J. F. Donges, I. Fetzer, S. J. Lade, M. Scheffer, R. Winkelmann, and H. J. Schellnhuber, Trajectories of the earth system in the anthropocene, Proc. Natl. Acad. Sci. USA **115**, 8252 (2018).

[9] Y. Cai, T. M. Lenton, and T. S. Lontzek, Risk of multiple interacting tipping points should encourage rapid CO_2 emission reduction, Nat. Clim. Chang. **6**, 520 (2016).

[10] T. M. Lenton, Environmental tipping points, Annu. Rev. Environ. Resour. **38**, 1 (2013).

[11] D. G. Wright and T. F. Stocker, A zonally averaged ocean model for the thermohaline circulation. I: Model development and flow dynamics, J. Phys. Oceanogr. **21**, 1713 (1991).

[12] A. Levermann and R. Winkelmann, A simple equation for the melt elevation feedback of ice sheets, Cryosphere **10**, 1799 (2016).

[13] A. Staal, E. H. van Nes, S. Hantson, M. Holmgren, S. C. Dekker, S. Pueyo, C. Xu, and M. Scheffer, Resilience of tropical tree cover: The roles of climate, fire, and herbivory, Glob. Chang. Biol. **24**, 5096 (2018).

[14] R. Abraham, A. Keith, M. Koebbe, and G. Mayer-Kress, Computational unfolding of double-cusp models of opinion formation, Int. J. Bifurc. Chaos **01**, 417 (1991).

[15] A. K. Klose, V. Karle, R. Winkelmann, and J. F. Donges, Dynamic emergence of domino effects in systems of interacting tipping elements in ecology and climate, arXiv:1910.12042.

[16] M. M. Dekker, A. S. von der Heydt, and H. A. Dijkstra, Cascading transitions in the climate system, Earth Syst. Dynam. **9**, 1243 (2018).

[17] D. J. Watts, A simple model of global cascades on random networks, Proc. Natl. Acad. Sci. USA **99**, 5766 (2002).

[18] Y. Yang, T. Nishikawa, and A. E. Motter, Small vulnerable sets determine large network cascades in power grids, Science **358**, eaan3184 (2017).

[19] B. Schäfer, D. Witthaut, M. Timme, and V. Latora, Dynamically induced cascading failures in power grids, Nat. Commun. **9**, 1975 (2018).

[20] H. Stommel, Thermohaline convection with two stable regimes of flow, Tellus A **13**, 224 (1961).

[21] Y.-H. Eom, Resilience of networks to environmental stress: From regular to random networks, Phys. Rev. E **97**, 042313 (2018).

[22] B. Bollobas, in *Random Graphs*, edited by W. Fulton, A. Katok, F. Kirwan, P. Sarnak, B. Simon, and B. Totaro (Cambridge University Press, Cambridge, UK, 2001).

[23] D. J. Watts and S. H. Strogatz, Collective dynamics of small-world networks, Nature **393**, 440 (1998).

[24] A. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).

[25] S. Milgram, The small-world problem, Psychol. Today **2**, 60 (1967).

[26] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Random graphs with arbitrary degree distributions and their applications, Phys. Rev. E **64**, 026118 (2001).

[27] P. W. Holland, K. B. Laskey, and S. Leinhardt, Stochastic block models: First steps, Soc. Networks **5**, 109 (1983).

[28] T. E. Oliphant, Python for scientific computing, Comput. Sci. Eng. **9**, 10 (2007).

[29] G. Fagiolo, Clustering in complex directed networks, Phys. Rev. E **76**, 026107 (2007).

[30] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Network motifs: Simple building blocks of complex networks, Science **298**, 824 (2002).

[31] M. Wiedermann, J. F. Donges, J. Kurths, and R. V. Donner, Spatial network surrogates for disentangling complex system structure from spatial embedding of nodes, Phys. Rev. E **93**, 042308 (2016).

[32] M. Hirota, M. Holmgren, E. H van Nes, and M. Scheffer, Global resilience of tropical forest and savanna to critical transitions, Science **334**, 232 (2011).

[33] A. C. Staver, S. Archibald, and S. A. Levin, The global extent and determinants of savanna and forest as alternative biome states, Science **334**, 230 (2011).

[34] C. Xu, S. Hantson, M. Holmgren, E. H. van Nes, A. Staal, and M. Scheffer, Remotely sensed canopy height reveals three pantropical ecosystem states, Ecology **97**, 2518 (2016).

[35] C. Ciemer, N. Boers, M. Hirota, J. Kurths, F. Müller-Hansen, R. S. Oliveira, and R. Winkelmann, Higher resilience to climatic disturbances in tropical vegetation exposed to more variable rainfall, Nat. Geosci. **12**, 174 (2019).

[36] E. H. van Nes, M. Hirota, M. Holmgren, and M. Scheffer, Tipping points in tropical tree cover: Linking theory to data, Glob. Chang. Biol. **20**, 1016 (2014).

[37] P. M. Cox, R. A. Betts, M. Collins, P. P. Harris, C. Huntingford, and C. D. Jones, Amazonian forest dieback under climate-carbon cycle projections for the 21st century, Theor. Appl. Climatol. **78**, 137 (2004).

[38] D. C. Zemp, H. M. J. Schleussner, Barbosa, M. Hirota, V. Montade, G. Sampaio, A. Staal, L. Wang-Erlandsson, and A. Rammig, Self-amplified amazon forest loss due to vegetation-atmosphere feedbacks, Nat. Commun. **8**, 14681 (2017).

[39] A. Staal, O. A. Tuinenburg, J. H. C. Bosmans, M. Holmgren, E. H. van Nes, M. Scheffer, D. C. Zemp, and S. C. Dekker, Forest-rainfall cascades buffer against drought across the Amazon, Nat. Clim. Chang. **8**, 539 (2018).

[40] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).

[41] J. C. Rocha, G. Peterson, Ö. Bodin, and S. Levin, Cascading regime shifts within and across scales, Science **362**, 1379 (2018).

Reprinted paper with permission from J. Krönke et al., Dynamics of tipping cascades on complex networks, Phys. Rev. E, vol. 101, no. 4, pp. 1-19, 2020, doi: 10.1103/PhysRevE.101.042311. Copyright 2020 by the American Physical Society

4
Analyses and studies of concrete cases and contexts

This fourth section moves from rather conceptual to more realistic real-world dynamics. We present analyses of concrete cases and contexts which create an empirical complement for our theoretical work.

A distinct understanding of social-ecological interactions is crucial to represent more realistic dynamics in World-Earth models. The subsection "Detecting complex social-ecological interactions in empirical data" (Sect. 4.1) presents our investigations and method developments on interaction dynamics in large-scale real-world data.

Interaction dynamics and social tipping are general concepts which occur in various areas. In the subsection on "Special cases of socio-economic dynamics and social tipping" (Sect. 4.2), the selected papers present specific nonlinear dynamics and positive feedback mechanisms emerging in complex socio-economic systems that are relevant for overall Earth system dynamics and sustainability.

In the last subsection, "Earth system analysis and planetary boundary interactions" (Sect. 4.3), we focus on coevolutionary interactions of attempts of Earth system stewardship and biogeophysical Earth system dynamics within the planetary boundaries.

4.1 Detecting complex social-ecological interactions in empirical data

In this first section we present investigations of interaction dynamics in large-scale real-world data.

We begin with a commentary paper on "Socio-economic data for global environmental change research" [Otto et al., 2015] in which we emphasize the need for subnational socio-economic datasets. Their existence is critical to assess the impacts of global environmental change and improving adaptation responses.

In particular, more data is needed on humanity's biggest emitters: the super-rich. In "Shift the focus from the super-poor to the super-rich" [Otto et al., 2019], we commented on this issue. In addition, we argued that more carbon mitigation policies are needed targeting the super-rich.

We continue with an examination of a prominent example of social-ecological interactions: the potential exacerbation of armed conflict by anthropogenic climate change and, in particular, by climate-related natural disasters. In "Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries" [Schleussner et al., 2016], we applied an event-coincidence analysis [Donges et al., 2016] based on data on armed conflict outbreaks and climate-related natural disasters.

We conclude with a complementary methodology, discussed in "Dose-response function approach for detecting spreading processes in temporal network data" [Donges, J. F. and Lochner, J. et al., 2021]. This work puts forward a methodology for the analysis of contagion dynamics in temporal complex networks based on dose-response functions and hypothesis testing using surrogate data sets.

been made. While such flexibility helped secure participation from all countries, the lack of detailed emissions information is problematic for understanding the impact of INDCs towards meeting global climate goals. Without this information, determining overlap between national, non-state, and subnational actions could become more difficult. Vague metrics may also provide cover for low ambition. Early analysis of the INDCs shows that current pledges are only half of what is needed to limit global temperature rise to 2 °C (ref. 16).

Leave room for innovation. At the same time, the criteria for inclusion should not be too strict. Some proponents argue for the integration of subnational and non-state actions into the UNFCCC. Others caution that this integration would prevent innovation and risk-taking among new actors. A major contributor to the Summit's success in engaging a diversity of participants was the flexibility afforded to the content of commitments. The Summit's openness brought in businesses and other actors who would have been otherwise hesitant to commit at such a high-level forum. Meetings like this could play a key role in fostering new thinking and ideas for addressing climate change, as they have lower costs of failure than a formal process such as the UNFCCC. Any framework that includes non-state and subnational participants must achieve a delicate balance between establishing a bar that boosts ambition but is not so high as to deter critical actors from joining.

States are no longer the only actors tackling climate change. The Summit represents a new mode of elevating the groundswell of non-state and subnational action into official political channels. This integration is crucial to making a fragmented climate governance system effective. Tenuous financing and uncertain implementation, however, mean that the Summit's commitments have a high risk of failure, potentially damaging the credibility of future non-state and subnational efforts. To avoid such a pessimistic conclusion, new methods of pledging and accountability, as well as innovative modes of governance, are needed to seriously engage new actors. ❐

Angel Hsu[1], Andrew S. Moffat[2], Amy J. Weinfurter[2] and Jason D. Schwartz[2] are at [1]Yale School of Forestry and Environmental Studies, Yale University, 195 Prospect Street, New Haven, Connecticut 06511, USA. [2]Yale Center for Environmental Law & Policy, Yale University, 195 Prospect Street, New Haven, Connecticut 06511, USA.

References

1. *Lima Call for Climate Action* Decision-/CP20 (UNFCCC, 2014); http://go.nature.com/ZRE3zU
2. Blok, K., Höhne, N., van der Leun, K. & Harrison, N. *Nature Clim. Change* **2,** 471–474 (2012).
3. *The Emissions Gap Report 2014* (UNEP, 2014); http://go.nature.com/lA6naE
4. Keohane, R. O. & Victor, D. G. *Perspect. Polit.* **9,** 7–23 (2011).
5. Biermann, F., Chan, S., Mert, A. & Pattberg, P. in *Public–Private Partnerships for Sustainable Development: Emergence, Influence and Legitimacy* (eds Pattberg, P. *et al.*) 69–87 (Edward Elgar, 2012).
6. Biermann, F., Pattberg, P., Van Asselt, H. & Zelli, F. *Glob. Environ. Polit.* **9,** 14–40 (2009).
7. Hale, T. & Mauzerall, D. *J. Environ. Dev.* **13,** 220–239 (2004).
8. Van Asselt, H. *The Fragmentation of Global Climate Governance: Consequences and Management of Regime Interactions* (Edward Elgar, 2014).
9. Widerberg, O. & Pattberg, P. *Global Policy* **6,** 45–56 (2014).
10. Chan, S. & Pauw, P. *A Global Framework for Climate Action (GCFA): Orchestrating Non-State and Sub-national Initiatives for More Effective Global Climate Governance* Discussion Paper 34 (German Development Institute, 2014).
11. Skocpol, T. *Naming the Problem: What It Will Take To Counter Extremism and Engage Americans in the Fight Against Global Warming* (Harvard Univ., 2013).
12. UN Climate Summit: Ban Ki-moon Final Summary. *UNFCCC* (September 25 2014); http://go.nature.com/25KTrU
13. *CO_2 Emissions From Fuel Combustion: Highlights* (IEA, 2013); http://go.nature.com/9DPilY
14. *Utility-Scale Energy Technology Capacity Factors* (NREL, 2014); http://www.nrel.gov/analysis/tech_cap_factor.html
15. *Rio+20 Voluntary Commitments* (UNCSD, 2012); http://go.nature.com/l53mth
17. Wolosin, M. & Belenky, M. *Gap Analysis with Paris Pledges* (Climate Advisors, 2014); http://go.nature.com/42InFe
18. IPCC *Climate Change 2013: The Physical Science Basis* (eds Stocker, T. F. *et al.*) (Cambridge Univ. Press, 2013).

Acknowledgements

We thank A. Zomer, P. Hirsch, and F. Melkinov of Yale University for providing research support. We are grateful to the participants within the Galvanizing the Groundswell network convened by the Natural Resources Defense Council, the Blavatnik School of Oxford University, and the Stanley Foundation for feedback.

Additional information

Supplementary information is available in the online version of the paper.

COMMENTARY:

Socio-economic data for global environmental change research

Ilona M. Otto, Anne Biewald, Dim Coumou, Georg Feulner, Claudia Köhler, Thomas Nocke, Anders Blok, Albert Gröber, Sabine Selchow, David Tyfield, Ingrid Volkmer, Hans Joachim Schellnhuber and Ulrich Beck

Subnational socio-economic datasets are required if we are to assess the impacts of global environmental changes and to improve adaptation responses. Institutional and community efforts should concentrate on standardization of data collection methodologies, free public access, and geo-referencing.

There is a scalar mismatch between social scientists focusing on the nation-state and climate scientists operating at the global level[1]. From the natural science perspective, climate change is an egalitarian and cross-border phenomenon, and research results are routinely analysed beyond national borders. The social sciences, however, have evolved historically within nation-states, and the production of data is mostly framed according to nation-state boundaries; this includes international comparisons. Overcoming this 'methodological nationalism' requires both cosmopolitan and subnational data[2].

Cosmopolitan data are needed to grasp the interconnectivity and interdependence of global, national and local issues. To obtain data at a subnational scale, for example on water use in different sectors

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Socio-economic data for global environmental change research, I. M. Otto et al., doi: 10.1038/nclimate2593, Copyright 2015, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.

and water prices, scientists usually have to visit the region and literally photocopy the information from local administrative organizations[3]. Such a process is time-consuming; also, data pooled from different countries and administrative units often use different methodologies and definitions and therefore must be standardized before use[4]. In contrast, the impacts of global environmental changes occur within climatological and geo-ecological units rather than administrative boundaries. Thus, the social impacts of global environmental changes may not be detectable by studying national averages.

In an illustration of this problem, we compare national and spatially explicit hunger indicators, and show that hunger is not equally distributed within national borders but is spatially concentrated in certain areas (Fig. 1). In many such areas, such as the Chad Lake Basin on the borders of Niger, Nigeria, Chad and Cameroon, for example, food production is threatened by decreasing and uncertain water availability[5]. The local effect of droughts on hunger occurrence or any other climate-induced socio-economic trend visible at the river-basin level is likely to disappear in averages at national level. At least 261 of the world's major rivers are shared, with 176 flowing through two countries, 48 through three countries, and 37 through four or more countries[6]. Although there are several programmes designed to exchange data within river basins, these primarily focus on hydrological data rather than socio-economic data[7].

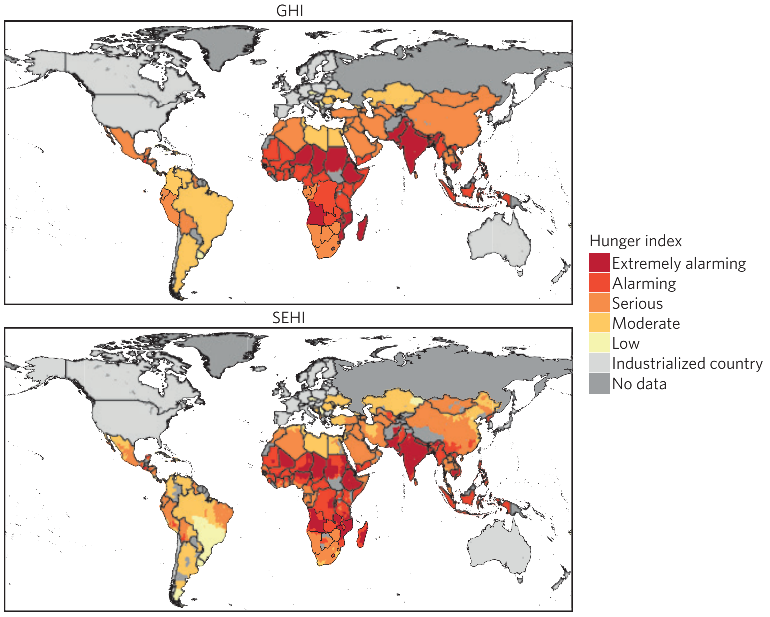


Figure 1 | A comparison of the Global Hunger Index (GHI) and the Spatially Explicit Hunger Index (SEHI). The global hunger index (GHI) provided by the International Food Policy Research Institute (IFPRI)[6] combines three equally weighted indicators: a national average of the proportion of people that are undernourished, and two subnational indicators — the percentage of underweight children younger than five, and the mortality rate of children younger than five[7,8]. In our spatially explicit hunger index (SEHI), we replace the national average of the proportion of people that are undernourished used by IFPRI by a subnational (0.5°) indicator, provided by the Food and Agriculture Organization[9], on the prevalence of stunting among children under five. The SEHI reveals that patterns of hunger are not bounded by national borders. The data are assembled for varying years from 2000 to 2011.

Stationarity in social sciences

To assess climate impacts and to develop strategies for adaptation and other global challenges, a different approach to data gathering and management is needed. Currently, most resources, externalities of economic activities and populations are not restricted to national borders; they become increasingly interconnected, and large and rapid shifts in these factors may occur. As an example, annually more people are reported to be displaced by natural disasters than by conflicts. By 2050, between 25 million and one billion people are projected to be forced to migrate because of climate change and other environmental factors[8]. Such estimates are mostly based on the physical occurrence of natural disasters, on which data exist. But there is no systematic database on current environmentally induced cross-border migration, nor on the number of people displaced by slowly occurring environmental changes[8], and no data on transit migration. The stationarity of data gathering has to be overcome[9,10] in social sciences, and the changes in our societies need to be reflected by use of new methods and new categories in socio-economic statistics.

In natural climate science, this process was initiated with the establishment of the International Meteorological Organization in the 1870s, succeeded by the World Meteorological Organization in 1951. These organizations instigated the consolidation and exchange of national weather data. It took many decades, however, to overcome national (military and commercial) interests and the inertia of installed infrastructure, and to standardize meteorological data on a global scale[11]. In fact, it is only since the start of the satellite age in the 1960s that an infrastructure for generating global weather and climate data has emerged.Today, climate scientists have access to snapshots of the state of the atmosphere every 6 hours, real-time information on the extent of Arctic sea-ice, continuously updated global data, and much more. They can also make use of records of temperature and precipitation that stretch as far back as the late nineteenth century, with near-global coverage. These datasets have proved invaluable for our understanding of climate change and of the role of natural variability and anthropogenic forcing, including attribution of extreme weather events[12,13]. Furthermore, they have triggered global 'system thinking', both in and outside the scientific community, highlighting the limits to our planetary resources.

A new paradigm in data gathering

Data and information to aid in the understanding of complex problems are key to the successful governance of common pool resources, including global commons[14]. To address urgent questions related to the world's foremost challenges, the social sciences and institutions gathering data will have to react and adapt more quickly to global challenges. Given current information and communication technologies, including the Internet, crowd sourcing and geographical information systems, and the fact that most national datasets are already digitized, this should be technically possible in a relatively modest time span. Available global geo-referenced databases, for example on demographic and economic indicators,

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Socio-economic data for global environmental change research, I. M. Otto et al., doi: 10.1038/nclimate2593, Copyright 2015, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.

Table 1 | Examples of existing global data sources relevant for researching social impacts of global environmental changes.

Indicator	Source	Lowest resolution level	Available years of observations
General demography			
Population density	Center for International Earth Science Information Network (CIESIN)	2.5′ × 2.5′ grid	2005, 2010, 2015
Population number, mortality, fertility	UN Population Division	National	1949–2012
Life expectancy	WHO, OECD, World Bank	National	1960–2012
Infant mortality rate	CIESIN	Subnational, 0.25° × 0.25° grid	2000
Education			
Literacy, school enrolment (by gender and age)	UN Gender Statistics	National	1990–2010, many missing observations
School enrolment	World Bank	National	1970–2012, many missing observations
Economic			
GDP per capita	Geographically Based Economic Database (G-Econ)	Subnational, 1° × 1° grid	2005
Food price	Food and Agriculture Organization	For several countries subnational at the province level, otherwise national	Monthly 2000 to present
Migration			
Persons of concern for UNHRC[a]	United Nations Human Rights Council (UNHRC)	Subnational at the province level	2000–2012, many missing observations
Asylum-seekers	UNHRC	National	2000–2012
International migrant stock	UN Population Division	National	1990–2010
Poverty			
Poverty rates in different age groups	OECD	National	1983–2011, many missing observations
Percentage of the population living on less than US$2.00 a day	WB	National	1980–2012, many missing observations
Child malnutrition	CIESIN	2.5′ × 2.5′ grid	2005
Behaviour and perceptions			
Perceived seriousness of global warming	World Value Survey	National	2009
Ecological footprint	Global Footprint Network	National	1961–2007

[a]Persons of concern for UNHRC including refugees, asylum-seekers, returned refugees, internally displaced persons (IDP), returned IDPs, stateless persons and others.

show that such efforts are possible (see Table 1). But these databases are only available for restricted time periods, and the highest spatial resolution available is usually the national level, often with many missing countries or ambiguous values. For example, the OECD and World Bank report different life expectancy values for the same countries over the same time period.

Homogenization of data collection methodologies, free public access to data at a subnational scale, together with geo-referencing of socio-economic data should be given the highest priority. Existing international organizations dealing with global environmental and social challenges could take a lead in this process.

Currently most international socio-economic data is collected by the United Nations Statistics Division (UNSD), to which data are supplied by National Statistics Offices through UNSD questionnaires and censuses[15]. The United Nations provides mandates to other international organizations such as the World Bank or the World Health Organization (WHO) to deal with specific data challenges such as on poverty or health. One possible move towards improving the subnational data accessibility would be to ask national statistical offices to add subnational entries in the UNSD questionnaires. The subnational level agreed on would have to be large enough to protect the anonymity of respondents, yet explicit enough to enable the disaggregation of national data. For example, the spatial resolution of 0.5° that corresponds to an area of 50 km^2 at the Equator (and is roughly the area administered by local governments in many modern nations) could be a suitable solution.

A short unpublished survey that we carried out among employees in statistic divisions of international organizations highlighted that implementing the above changes would require more data scientists and different data management strategies. It was also pointed out that providing homogenized subnational-level data, especially in low-income countries, would require substantial improvements to the local data collection infrastructure. These are important challenges that would have to be overcome by international agreements and the reallocation of funding necessary for improving data infrastructure and management.

In addition, bottom-up and crowd data pooling initiatives should be encouraged. There are numerous regional case studies and research involving household surveys being carried out all over the world, and good scientific practice codes could encourage standardization of data gathering and data accessibility. Improved information exchange and information access can help to generate a better understanding and awareness of the interconnectedness between global environmental changes and social impacts, and through this, increased adaptation capacity at the global and local levels. ❐

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Socio-economic data for global environmental change research, I. M. Otto et al., doi: 10.1038/nclimate2593, Copyright 2015, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.

Ilona M. Otto[1,2], Anne Biewald[1], Dim Coumou[1], Georg Feulner[1], Claudia Köhler[1], Thomas Nocke[1], Anders Blok[3], Albert Gröber[4], Sabine Selchow[4,5], David Tyfield[4,6], Ingrid Volkmer[7], Hans Joachim Schellnhuber[1,8] and Ulrich Beck[4] are at: [1]Potsdam Institute for Climate Impact Research, Telegraphenberg A31, Potsdam, 14473, Germany; [2]Zhejiang University, School of Public Affairs, Yuhangtang Road 866, Hangzhou 310,058, China; [3]University of Copenhagen, Department of Sociology, Øster Farimagsgade 5, Postboks 2099, Copenhagen 1014, Denmark; [4]Ludwig-Maximilians-Universität, Institute for Cosmopolitan Studies, Konradstrasse 6/203, Munich 80801, Germany; [5]London School of Economics and Political Science, Department of International Development, Houghton Street, London WC2A 2AE, UK; [6]Lancaster University, Lancaster Environment Centre, Bailrigg, Lancaster, LA1 4YT, UK; [7]University of Melbourne, School of Culture and Communication, Victoria 3,010, Australia; [8]Santa Fe Institute, Santa Fe, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.*

**e-mail: ilona.otto@pik-potsdam.de*

References

1. Bakker, K. *Science* **337,** 23–24 (2012).
2. Beck, U. & Grande, E. *Br. J. Sociol.* **61,** 409–443 (2010).
3. Wang, X., Otto, I. M. & Yu, L. *Agric. Water Manag.* **119,** 10–18 (2013).
4. Montgomery, M. R. *Science* **319,** 761–764 (2008).
5. Onuoha, F. C. *Afr. J. Conflict Resolut.* **8,** 35–61 (2008).
6. Myers, N. in *Conf. Pap. The Hague Conf. Environ. Secur. Sust. Dev.* (Institute for Environmental Security, 2004); http://www.envirosecurity.org/conference/working/newanddifferent.pdf
7. Gerlak, A. K., Lautze, J. & Giordano, M. *Int. Environ. Agreements Polit. Law Econ.* **11,** 179–199 (2010).
8. UNHCR *The State of the World's Refugees. In Search of Solidarity* (Oxford Univ. Press, 2012).
9. Beck, U. *Glob. Netw.* **2,** 165–181 (2010).
10. Kundzewicz, Z. W. *et al. Hydrol. Sci. J.* **53,** 37–41 (2008).
11. Edwards, P. N. *A Vast Machine: Computer Models,Climate Data, and the Politics of Global Warming* (MIT Press, 2013).
12. IPCC *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation* (eds Field, C. B. *et al.*) (Cambridge Univ. Press, 2012).
13. Coumou, D., Robinson, A. & Rahmstorf, S. *Climatic Change* **118,** 771–782 (2013).
14. Ostrom, E. *Glob. Environ. Change* **20,** 550–557 (2010).
15. *Major Work Areas and Accomplishments* (UNSD, 2014); http://go.nature.com/BFpHG1

Acknowledgements

We dedicate this paper to the memory of Professor Ulrich Beck (1944–2015), the pioneer of cosmopolitan social science in a world of global climate risks.

COMMENTARY:

Local science and media engagement on climate change

Candice Howarth and Richard Black

Climate scientists can do a better job of communicating their work to local communities and reignite interest in the issue. Local media outlets provide a unique opportunity to build a platform for scientists to tell their stories and engage in a dialogue with people currently outside the 'climate bubble'.

Surveys, including those carried out regularly by the UK's Department of Energy and Climate Change (DECC), show that a majority of the British public accept that climate change is happening, are concerned about it, and favour action to reduce greenhouse-gas emissions[1]. However, public acceptance of climate change has reduced over the past five years. This may be connected with a lack of appreciation of the scientific consensus, which by several measures exceeds 90% (ref. 2). In 2014, a ComRes survey of 2,000 members of the British public, commissioned by the Energy and Climate Intelligence Unit, found that only 11% of respondents appreciated the extent of the scientific consensus on climate change; nearly half (47%) did not think there was a consensus at all[3]. Although the DECC (and other) surveys regularly show high levels of support for renewable energy technologies such as wind and solar power, the ComRes survey found that only 5% of the population knows that support is this high; more than half of the population (63%) thinks that the public is opposed.

The methods by which people receive, interpret and understand information on climate change is important as it affects their resulting actions[4]. The importance and relevance of place attachments in understanding human responses to climate change is known[5], and by incorporating elements of 'daily life' (which by definition is lived at a local level), media portrayals can enable climate science and governance to be interpreted through a local, everyday lens[6].

Yet the communication of climate change historically has been generic, untailored and untargeted. A transition to a situation in which public engagement on climate change goes beyond information provision and instead adopts a more active approach underpinned by constructive dialogue between scientists and the media could therefore be fruitful. Increasing engagement on the local dimensions of climate change could facilitate this and enable a stronger connection to the issue.

The 2013–2014 winter saw a sequence of serious flooding events across much of the UK. Both a survey commissioned by Avaaz at the height of the floods[7] and the ComRes survey six months later, suggested that these events affected public opinion on climate change. In the first, nearly half of respondents said they believed the floods were linked to climate change. In the second, half said that the floods had increased their belief in climate change, and a quarter said it increased their belief in human agency. The flooding was a major story on national and regional media for weeks and the subject of intense political discourse, and these studies could not untangle the question of whether local or national factors were involved in people making the weather-climate link. However, a study on the 2012 floods in Wales[8] indicated that local experience is important; people directly exposed to flooding were more likely to accept evidence for climate change, and to believe that their own actions could have an impact by reducing carbon emissions.

comment

Shift the focus from the super-poor to the super-rich

Carbon mitigation efforts often focus on the world's poorest people, dealing with topics such as food and energy security, and increased emissions potential from projected population, income and consumption growth. However, more policies are needed that target people at the opposite end of the social ladder — the super-rich.

Ilona M. Otto, Kyoung Mi Kim, Nika Dubrovsky and Wolfgang Lucht

In 2017, there were just over 36 million adults classified as High Net Worth Individuals (net assets above US$1 million), and there were 148,000 classified as Ultra High Net Worth Individuals (net assets above US$50 million)[1]. The super-rich are, on the one hand, the most visible social group in terms of their presence in mass culture, social media, politics and business, and on the other hand, the most hidden social group in terms of the availability of data on their income, lifestyles, resource use, consumption patterns, mobility and social networks. It seems as though we know a lot about them from watching television and soap operas, and reading glossy magazines.

However, once we try to obtain more concrete data about this social group, there is practically nothing available and in practice very few people personally know someone belonging to the super-rich. For example, the supposedly representative survey of the German population on per capita consumptions of natural resources largely omits the most-wealthy respondents; it includes only 3.5% of respondents that reported income above €5,000 per month (ref. [2]). According to the German Statistical Office, however, 15.1% of households in Germany have a monthly income in the range €5,000–18,000 (ref. [3]).

Affluent people can more easily disconnect themselves from the realities of climate change and climate extremes[4], and are in general the least affected by natural disasters, against which they can shield themselves more effectively; their extreme mobility gives them options to avoid dangerous environmental situations and they have greater economic capacity and better accessibility to recovery systems. This perhaps explains why the most wealthy have been largely ignored in climate change research, which instead frequently focuses on the poor, who are the group most affected by, and most vulnerable to, climate change impacts.

However, given their notable affluence in lifestyle and consumption when compared to the poor, a better understanding of the super-rich could be an important contribution to climate mitigation options. The lifestyles and consumption patterns of the super-rich strongly influence the globally growing middle classes, who emulate upper-class consumption styles to distinguish themselves from lower classes[5]. In addition, the super-rich have a great impact on technological innovation and could actively support zero carbon and renewable energy technologies. The world's billionaires have driven almost 80% of the 40 main breakthrough innovations over the last 40 years (ref. [6]). Moreover, consumption choices of the wealthiest could support market penetration of new technologies that are still not affordable for the middle classes.

Here we estimate the greenhouse gas emissions of the super-rich to suggest the carbon savings that could be obtained by targeting this group, and we reflect on how this could be achieved.

Emissions of the super-rich

There are just a few scientific publications analysing lifestyles and associated greenhouse gas emissions of the super-rich, that is, their personal lifestyle emissions rather than those of the investment assets they may additionally hold or control as a part of their wealth, and none based on representative surveys. According to some estimates, the average lifestyle consumption carbon footprint of someone in the richest 1% could be 175 times that of someone in the poorest 10% (ref. [7]).

We conducted lifestyle consumption surveys with four interviewees including three super-rich people and a pilot operating a private jet that is hired by private wealthy customers. From this data, we have averaged the results from four online carbon-footprint calculators to estimate the carbon emissions corresponding to the lifestyles reported by our interviewees (see Table 1).

The households that we interviewed are each believed to hold over US$1 million in investment assets excluding their primary residence and personal items; two families were living in South Korea and one in the United States. The pilot had customers primarily from Central Europe. He provided us with the average annual distance and number of flights of his customers. Our survey focused on emissions from private motor vehicles, air travel, household energy use and spending on food and education. These activities arguably cover about 70–80% of carbon emissions from individual consumption[8].

Our results suggest that a typical super-rich household of two people produces a carbon footprint of 129.3 tCO_2e per year. Motor vehicle use generates approximately 9.6 tCO_2e per year, with household energy emitting 18.9 tCO_2e per year, secondary consumption 34.3 tCO_2e per year, and 66.5 tCO_2e per year generated by the leading emission contributor: air travel (Fig. 1). Our carbon emissions estimates are substantially lower than those provided by Chancel and Piketty[9] in an analysis based on national GDP and emission data for the years 1998–2013, but amount to around ten times that of the global per-person average. Calculating the emissions from 0.54% of the wealthiest of the global population, according to our estimates, results in cumulative emissions equal to 3.9 billion tCO_2e per year. This is equivalent to 13.6% of total lifestyle-related carbon emissions. In comparison, the world's poorest 50% are responsible for about 10% of lifestyle consumption emissions[7].

Room for reduction

There is a largely untapped potential to reduce carbon emissions by altering the way of life of the super-rich. For example, reducing the carbon footprint of this group by about 20% could be achieved by turning their residences into zero carbon homes with decentralized renewable energy production and using electric vehicles for

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Shift the focus from the super-poor to the super-rich, I. M. Otto et al., doi: 10.1038/s41558-019-0402-3, Copyright 2019, Springer Nature Limited.

Table 1 | Summary of survey data collected on the monthly consumption habits reported by four interviewees. The averaged results from four different carbon-footprint calculators were used to estimate the emissions of a typical super-rich household.

	Interviewee A	Interviewee B	Interviewee C	Interviewee D
Business sector	Investment Real estate	Trade	Aviation	Investment Finance
Household size	1	1	–	5 (and 2 babysitters)
Motor vehicles	2 Discovery Sport Mercedes E Coupe	3 Mercedes C63 GranTurismo Hyundai Genesis (excluded from data)	–	2 Large sedans
Driving miles[a]	Discovery Sport: 800 miles Mercedes E Coupe: 400 miles	Mercedes C63: 1,000 miles GranTurismo: 600 miles	-	Car 1: 1,564 miles Car 2: 1,279 miles
Air travel[b]	Short: 5 Medium: 2 Long: 1	Short: 0 Medium: 10 Long: 0	Total distance: 4,143 miles	Short: 0 Medium: 0 Long: 2
Houses	2 (Republic of Korea) 1: 280m^2 2: 185m^2	2 (Republic of Korea and Thailand) 1: 185m^2 2: 560m^2	–	2 (United States) 1: 500m^2 2: 500m^2
Secondary	–	–	–	Cultural activities: US$2,000 Food: US$2,500 Education: US$2,083–5,000
Carbon footprint[c] (tCO$_2$e per year)	73.3	84.7	177.4	105.6

[a]Driving mileage expressed in miles for two most frequently used cars. [b]A one-way flight is counted as 1. [c]The average result of calculating the carbon footprint with four different carbon footprint calculators: CoolClimate Network (http://coolclimate.berkeley.edu/calculator); Carbon Footprint (https://www.carbonfootprint.com/calculator.aspx); myclimate (http://www.myclimate.org/); Korean Carbon Footprint (http://www.kcen.kr/tanso_20120314/main.html).

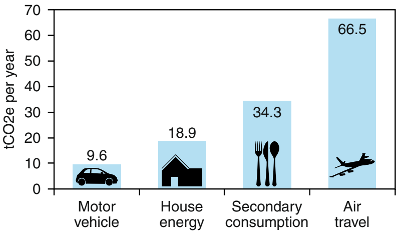


Fig. 1 | The estimated carbon footprint of a typical super-rich household of two people. Data were derived from four consumption habit surveys, and show the average of four carbon-footprint calculators for each of four consumption categories. Total emissions are approximately 129.3 tCO$_2$e per year.

both energy storage and land transport. Some secondary consumption emissions could be avoided by choosing more durable goods and reducing consumption. Frequent air travel is a primary contributor to hugely above-average emissions of the super-rich that could be substantially reduced by avoiding using private jets and just flying less. Changes in behaviour of the super-rich to reduce their emissions may also have important down-stream benefits, as their lifestyles are the sources of inspiration for the consumption behaviour of the rest of the population.

Some of the wealthiest people are known to already actively engage in climate protection. For example, Bill Gates supports and invests in combatting climate-change-related problems, through the Bill & Melinda Gates Foundation. Otto Group as well as the Bosch Company are associated with foundations that actively support environmental and sustainability-oriented research and education. Stordalen Foundation has invested in a wide range of cutting-edge research and public engagement for sustainability. Other super-rich have been planting trees in an effort to offset their carbon footprints[10]. Nevertheless, these examples are far from typical, and it is the unengaged majority of the super-rich that requires attention if substantial emissions reductions are to be achieved.

Policies must target the super-rich

The wealthiest are not much affected by the mitigation policies in which nation states are the main actors as well as the main sources of funding. The current climate mitigation efforts focusing on afforestation, energy supply and demand, transportation and buildings[11] correlate only weakly with the sectors driving the world's biggest fortunes (finance and investment, fashion and retail, and real estate[12]). Heavy environmental taxation, as commonly discussed, is unlikely to effect the consumption behaviour of the super-rich, who can afford to continue polluting[4].

Policies that more aggressively force carbon-footprint reduction of the super-rich may be pursued as a part of a comprehensive portfolio of mitigation. Examples of policies that are currently being discussed include compulsory restrictions on household and individual emissions, and building code regulations[13]. Those specifically targeting the wealthiest could include obligatory installation of renewable energy facilities on houses and apartments above a certain size. Importantly, in contrast to the poorest in the community, the richest have the agency and power needed to change their lifestyles to meet policy requirements without compromising quality of life. The leadership of the super-rich in adopting renewable energy technologies could generate positive knowledge and technology diffusion spill-over effects, making such technologies more attractive and more affordable for other social groups.

In addition, new and more sophisticated policy instruments are needed. Some authors propose introducing an inheritance tax[14,15] that could be an additional source of funds for climate mitigation. In 2017 alone, 44 heirs inherited more than a billion dollars each, totalling US$189 billion (ref. [6]). For comparison, the four largest multilateral climate funds, the Green Climate Fund,

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Shift the focus from the super-poor to the super-rich, I. M. Otto et al., doi: 10.1038/s41558-019-0402-3, Copyright 2019, Springer Nature Limited.

Adaptation Fund, Climate Investment Funds and Global Environment Facility, approved a total of US$2.78 billion of project support in 2016 (ref. [16]).

Next steps

Any form of policy targeted at the super-rich is bound to meet with strong resistance. The rich are over-represented in national governments and there are strong ties between the wealthy and the political elites. Therefore, it is important to raise awareness about these issues and to build social pressure on the super-rich and political elites all over the world.

More research is also needed to understand the motives that might drive the wealthy to become environmentally engaged in their private life as well as in their business operations. For example, major investors could be encouraged to exert influence on the fossil-fuel sector by divesting their assets and reinvesting their money in renewables, however, one would have to understand first which arguments and communication channels should be used to successfully reach this group.

Finally, more efforts are needed to educate the rich. The impacts of unmitigated climate change on ecosystems, agricultural production and water availability in the twenty-first century will lead to large-scale population displacements, disruption of international trade networks, food shortages and an increasing number of conflicts over basic resources[17]. The manifold consequences for human security and health suggest that no amount of money would guarantee the safety, or even survival, of our generation's offspring, including those from super-rich families. Such a message should reach the world's most wealthy and most powerful. ❐

Ilona M. Otto[1*], Kyoung Mi Kim[2,3], Nika Dubrovsky[5] and Wolfgang Lucht[1,4]

[1]Earth System Analysis, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, Potsdam, Germany. [2]Environmental and Resource Management, Brandenburg University of Technology Cottbus – Senftenberg, Cottbus, Germany. [3]Korea Foundation for the Advancement of Science and Creativity, Seoul, South Korea. [4]Department of Geography, Humboldt-Universität zu Berlin, Berlin, Germany. [5]Unaffiliated: dubrovsky@gmail.com.
**e-mail: ilona.otto@pik-potsdam.de*

Published online: 28 January 2019
https://doi.org/10.1038/s41558-019-0402-3

References

1. *Global Wealth Report 2017* (Credit Suisse AG, 2017).
2. Kleinhückelkotten, S., Neitzke, H.-P. & Moser, S. *Repräsentative Erhebung von Pro-Kopf- Verbräuchen natürlicher Ressourcen in Deutschland (nach Bevölkerungsgruppen)* 143 (2016).
3. *Sample Survey of Income and Expenditure* (Statistisches Bundesamt, 2016).
4. Kenner, D. *Inequality of Overconsumption: The Ecological Footprint of the Richest* (Anglia Ruskin University, 2015).
5. Kravets, O. & Sandikci, O. *J. Mark.* **78**, 125–140 (2014).
6. *Billionaires Insights 2018: New Visionaries and the Chinese Century* (USB, PwC, 2018).
7. Oxfam. *Extreme Carbon Inequality* (Oxfam, 2015).
8. Peattie, K. & Peattie, S. *J. Bus. Res.* **62**, 260–268 (2009).
9. Chancel, L. & Piketty, T. *Carbon and Inequality: from Kyoto to Paris* (2015).
10. Rapier, R. Leonardo DiCaprio's carbon footprint is much higher than he thinks. *Forbes* (1 March 2016).
11. IPCC *Climate Change 2014: Mitigation of Climate Change* Ch. 6 (eds Edenhofer, O. et al.) (Cambridge Univ. Press, 2014).
12. How the world's billionaires got so rich. *Forbes* (10 March 2018).
13. Keskitalo, E. C. H., Juhola, S., Baron, N., Fyhn, H. & Klein, J. *Climate* **4**, 7 (2016).
14. Puaschunder, J. M. *Mapping Climate Justice* (Social Science Research Network, 2016).
15. WBGU. *Just and In-Time Climate Policy Four Initiatives for a Fair Transformation* (German Advisory Council on Global Change, 2018).
16. Mapped: Where multilateral climate funds spend their money. *Carbon Brief* (6 November 2017).
17. Schellnhuber, H. J. et al. in *Handbook on Sustainability Transition and Sustainable Peace* 267–283 (Springer International Publishing, 2016).

Acknowledgements

The authors acknowledge comments on the research received from F. Wätzold, K. Vinke, J. Sendzimir, V. Hüzeir, T. Garcia and A. Schuster as well as funding from the Earth League alliance. This work has been carried out within the Copan: Co-evolutionary Pathways Research Group at PIK. N.D. is an independent artist and author of the international educational platform Anthropology For Kids (www.A4Kids.org).

Grounding nature-based climate solutions in sound biodiversity science

The current narrow focus on afforestation in climate policy runs the risk of compromising long-term carbon storage, human adaptation and efforts to preserve biodiversity. An emphasis on diverse, intact natural ecosystems — as opposed to fast-growing tree plantations — will help nations to deliver Paris Agreement goals and much more.

Nathalie Seddon, Beth Turner, Pam Berry, Alexandre Chausson and Cécile A. J. Girardin

The idea that natural ecosystems can help us fight both the drivers and impacts of climate change has been gaining traction over the past few years, including recent emphasis in the IPCC Special Report[1]. In particular, the Paris Agreement on climate change calls on all parties to acknowledge "the importance of ensuring the integrity of all ecosystems, including oceans, and the protection of biodiversity, recognized by some cultures as Mother Earth", and 66% of signatories to the agreement commit to 'green' or 'nature-based solutions' in their climate pledges (see Nature-Based Solutions Policy Platform; www.nbspolicyplatform.org) (Box 1). Such recognition of nature's value — in particular through policies promoting forests as carbon sinks — was hard-won by negotiators and non-state actors and is vitally important. However, we are concerned by aspects of the narrative reaching policymakers, and call on scientists studying biodiversity and ecosystem functions and services to fully engage with and inform the process by which high-level pledges are translated into on-the-ground actions.

A focus on forests

When it comes to high-level multilateral pledges for nature, the current focus is on forests. The Bonn Challenge — launched by the International Union for Conservation of Nature (IUCN) and Germany in 2011 and currently involving 56 nations — is a global effort to restore 150 million hectares of deforested and degraded land by 2020 and 350 million hectares by 2030[2]; the New York Declaration on Forests — signed in 2014 by 37 governments, 63 non-governmental organizations, 53 multinational companies

Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries

Carl-Friedrich Schleussner[a,b,c,1], Jonathan F. Donges[a,d], Reik V. Donner[a], and Hans Joachim Schellnhuber[a,e,1]

[a]Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany; [b]Climate Analytics, 10969 Berlin, Germany; [c]Integrative Research Institute on Transformations of Human–Environment Systems, Humboldt University, 10099 Berlin, Germany; [d]Stockholm Resilience Centre, Stockholm University, 114 19 Stockholm, Sweden; and [e]Santa Fe Institute, Santa Fe, NM 87501

Contributed by Hans Joachim Schellnhuber, May 20, 2016 (sent for review February 5, 2016; reviewed by Yoshito Hirata and Jürgen Scheffran)

Social and political tensions keep on fueling armed conflicts around the world. Although each conflict is the result of an individual context-specific mixture of interconnected factors, ethnicity appears to play a prominent and almost ubiquitous role in many of them. This overall state of affairs is likely to be exacerbated by anthropogenic climate change and in particular climate-related natural disasters. Ethnic divides might serve as predetermined conflict lines in case of rapidly emerging societal tensions arising from disruptive events like natural disasters. Here, we hypothesize that climate-related disaster occurrence enhances armed-conflict outbreak risk in ethnically fractionalized countries. Using event coincidence analysis, we test this hypothesis based on data on armed-conflict outbreaks and climate-related natural disasters for the period 1980–2010. Globally, we find a coincidence rate of 9% regarding armed-conflict outbreak and disaster occurrence such as heat waves or droughts. Our analysis also reveals that, during the period in question, about 23% of conflict outbreaks in ethnically highly fractionalized countries robustly coincide with climatic calamities. Although we do not report evidence that climate-related disasters act as direct triggers of armed conflicts, the disruptive nature of these events seems to play out in ethnically fractionalized societies in a particularly tragic way. This observation has important implications for future security policies as several of the world's most conflict-prone regions, including North and Central Africa as well as Central Asia, are both exceptionally vulnerable to anthropogenic climate change and characterized by deep ethnic divides.

climate-related natural disasters | ethnic fractionalization | armed conflicts | event coincidence analysis

Climate-related natural disasters are among the most important environmental stressors affecting the development of human societies. Climatic changes—and most prominently the succession of severe natural disasters—have been recognized as an important potential driver for the collapse of complex societies (1). However, not the climatological events per se, but societal vulnerability to its consequences in conjunction with other stressors has led to societal disintegration, armed conflicts, and eventually societal collapse during historic and prehistoric times (2–8). Today, armed conflicts are still among the biggest threats to human societies, and the identification of underlying processes and potential drivers is an area of intense scientific research. Several potential risk enhancement factors for conflict outbreak have been identified, including poverty (9), income inequality (10), weak governance (11), or a preexisting history of conflicts (12). Hypotheses relating to conflict feasibility based on financial assets from natural resource exploitation have also been discussed (13, 14). Additionally, there is a growing body of literature that reports robust indications that ethnic fractionalization is one of the key determinants of armed-conflict outbreak risk (10, 14–17). Although not necessarily rooting in ethnic tension, nearly two-thirds of all civil wars since 1946 have been fought along ethnic lines (18). This prominent role of ethnicity in conflicts might be related to selective access to political power or resources that are often divided along ethnic lines (19), as well as to a high and rapid ethnic mobilization potential (20) arising from geographical clustering of ethnic groups and strong interethnic social ties (21). These two factors may contribute to societal fissures along ethnic boundaries in case of rapidly emerging societal tension stemming from disruptive events such as natural disasters. In addition, it seems plausible that ethnic groups can be impacted very differently by natural disaster occurrence. The prevalent geographic clustering might be reinforced by other factors such as ethnically specific livelihoods (e.g., pastoral or riverine communities) or socioeconomic discrimination resulting in an ethnicity-dependent differential vulnerability to natural disasters (22).

In our analysis, we investigate the hypothesis that climate-related natural disasters (in the following referred to as disasters) enhance the risk of an emergence or violent outbreak of armed conflicts particularly in ethnically fractionalized societies. We explicitly address the impact of such disasters in terms of the resulting economic damage relative to national gross domestic product (GDP), making use of a high-quality database developed for commercial purposes of the reinsurance sector (*Materials and Methods*). Thereby, we explicitly define disasters with respect to their economic impact instead of the associated climatic variables. To test for statistical interrelationships between these damage events and the timing of armed conflicts, we use event coincidence analysis (ECA; see refs. 23 and 24, and Fig. 1 and *Materials and Methods*), a method that is conceptually related to event synchronization (25) and similar approaches that are widely used in the neurosciences for studying neuronal spike trains (26). ECA provides a generally applicable tool for explicitly testing the statistical significance of interdependences between sequences of events and has been proven useful in analyzing relations between

Significance

Ethnic divides play a major role in many armed conflicts around the world and might serve as predetermined conflict lines following rapidly emerging societal tensions arising from disruptive events like natural disasters. We find evidence in global datasets that risk of armed-conflict outbreak is enhanced by climate-related disaster occurrence in ethnically fractionalized countries. Although we find no indications that environmental disasters directly trigger armed conflicts, our results imply that disasters might act as a threat multiplier in several of the world's most conflict-prone regions.

Author contributions: C.-F.S., J.F.D., R.V.D., and H.J.S. designed research; C.-F.S. performed research; C.-F.S. and J.F.D. analyzed data; and C.-F.S., J.F.D., R.V.D., and H.J.S. wrote the paper.

Reviewers: Y.H., Institute of Industrial Science, University of Tokyo; and J.S., University of Hamburg.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data in this paper have been deposited at www.pik-potsdam.de/research/publications/pnas/Schleussner_et_al_2016_PNAS_scripts.zip.

[1]To whom correspondence may be addressed. Email: john@pik-potsdam.de or schleussner@pik-potsdam.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1601611113/-/DCSupplemental.

Reproduced from: C. F. Schleussner et al., Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries, Proc. Natl. Acad. Sci. U. S. A., vol. 113, no. 33, pp. 9216-9221, 2016, doi: 10.1073/pnas.1601611113.

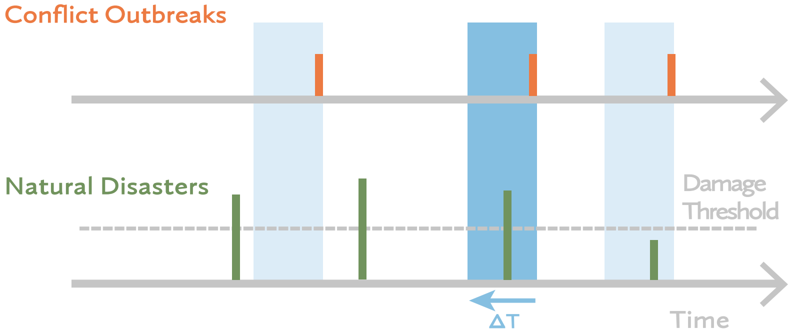


Fig. 1. Illustration of the methodological approach of event coincidence analysis for the risk enhancement test based on armed-conflict occurrence. An armed-conflict outbreak (orange) is counted as coincident with a natural disaster (green), if it co-occurs with or is preceded by such an event exceeding a prescribed damage threshold within a given coincidence interval ΔT.

event time series such as regime shifts in African paleoclimate and the appearance and disappearance of hominin species during the Plio-Pleistocene (23), plant growth response to climatic extremes (27, 28), or the role of flood events as triggers of epidemic outbreaks (24).

ECA allows to quantify the strength and robustness of statistical interrelationships between event series of natural disasters and armed-conflict outbreaks in two complementary ways (*Materials and Methods*): (*i*) the "risk enhancement test" is based on the "aggregated precursor coincidence rate" (24) measuring the fraction of conflicts that co-occurred with or were preceded by at least one disaster exceeding a certain damage level in the same country and that occurred at most at time ΔT before the conflict started (Fig. 1). In this case, a robust coincidence rate would indicate that disaster occurrence is a risk-enhancing factor for armed-conflict outbreak, based on a retrospective analysis with the condition that such an outbreak has occurred. (ii) In turn, the "trigger test" relies on the "aggregated trigger coincidence rate" (24) measuring the fraction of disasters exceeding a prescribed damage level in a country group that co-occurred with or were followed by at least one conflict that occurred at most a time ΔT after the disaster onset in the same country. This analysis allows to assess more explicitly than the risk enhancement test whether disasters may act as a direct trigger to armed-conflict outbreaks in the database under consideration. Statistical significance is tested with respect to an appropriately chosen null model (*Materials and Methods*), and we vary the economic damage threshold for identifying disasters to test for the effect of the event severity on the coincidence rate and significance as well as different disaster types (climatological, meteorological, and hydrological disasters; *SI Appendix*, Table S1).

Besides testing for a global relation between natural disaster occurrence and armed-conflict outbreak, we performed our analysis on a group of 50 countries with the highest ethnic fractionalization (EF) following a well-established ethnic fractionalization index (29) (results for different group sizes are given in *SI Appendix*). Additionally, we grouped countries according to alternative hypotheses such as multiple conflict outbreaks (CONFL, see ref. 12) and income inequality measured by the Gini coefficient (GINI, 50 countries with highest inequality; see Fig. 2 for the country classification). We furthermore analyzed other country groupings such as countries with high religious fractionalization, low levels of overall development, low literacy rates, abundant absolute poverty, high dependency on agricultural production, high corruption levels, or countries markedly affected by the El Niño Southern Oscillation (*SI Appendix*, Table S3). It is important to highlight that such a country grouping approach does not allow for a robust assessment of the relevance of different factors for the risk of armed-conflict outbreak generally, but rather indicates specific vulnerability to climate-related natural disaster impacts.

Results

In the following, we present the results of ECA of armed-conflict outbreaks listed in the UCDP/PRIO conflict dataset (30, 31) with natural disasters based on the NatCatSERVICE database from Munich Re over the period from 1980 to 2010 (32). We find no statistically significant precursor coincidence rates for the risk enhancement test at the global scale and all disaster types, except for the most devastating disasters that caused damage above 10% of annual country GDP (compare Fig. 3). As the database contains only about 40 events of this damage class globally (*SI Appendix*,

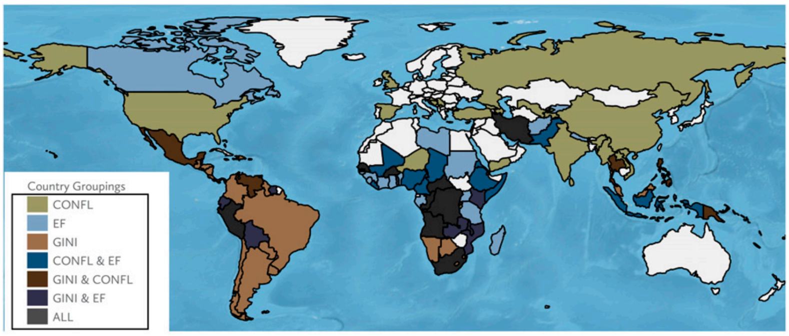


Fig. 2. Mapping of countries according to different analysis criteria including countries with more than one conflict (CONFL), the 50 countries with the highest Gini coefficient (GINI), as well as the 50 countries with the highest ethnic fractionalization (EF).

Reproduced from: C. F. Schleussner et al., Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries, Proc. Natl. Acad. Sci. U. S. A., vol. 113, no. 33, pp. 9216-9221, 2016, doi: 10.1073/pnas.1601611113.

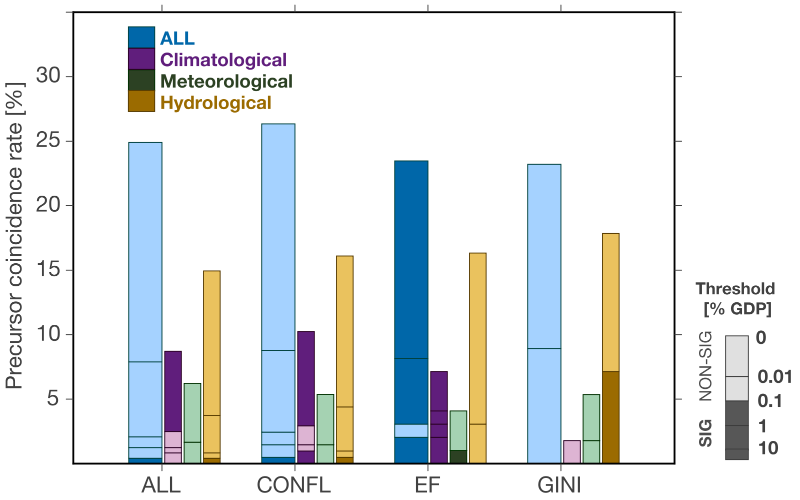


Fig. 3. Results of ECA for the risk enhancement test: the percentage of armed-conflict outbreaks that coincide with a climate-related natural disaster within the same month (*Materials and Methods*). We resolve different country groupings, disaster types (color coding), and disaster damage levels. Damage levels are indicated by segments of the individual bars and are assessed relative to the country's GDP in the year of the event. Segmenting starts with zero threshold from the top and the number of segments with nonzero coincidences can differ between country groupings and disaster types. Filled segments indicate coincidence rates that are significant at the 95% level. Results shown are for coincidences between events occurring within the same month (see *SI Appendix*, Fig. S3 for results for coincidence intervals of up to 12 months).

Table S2), however, no robust conclusions can be drawn for this category. To the contrary, our analysis reveals robust precursor coincidence within the same month for EF largely independent of the damage threshold resulting in a maximum coincidence rate of about 23% for all disaster types, which corresponds to 23 conflict outbreaks in total (see *SI Appendix*, Table S4, for an overview of the number of conflict outbreaks per country grouping). This finding is largely robust with regard to the arbitrarily chosen size of the country grouping (*SI Appendix*, Fig. S1). The results for the GINI, CONFL, as well as alternative country groupings (*SI Appendix*, Fig. S2) do not differ substantially from the global assessment. Despite existing linkages between some of the aforementioned factors and EF (compare Fig. 2), none of these country groupings yields results of similar robustness. In addition, we analyzed immediate and longer-term responses to disaster impacts (*SI Appendix*, Fig. S3). Although we find significant precursor coincidence rates for an extended coincidence interval of up to 3 months before the conflict outbreak, our analysis does not reveal significant effects for longer intervals.

A different picture emerges when different types of disasters are treated separately (see *SI Appendix*, Table S1, for further details on the event type classification). About 9% of all global armed-conflict outbreaks (21 in total) significantly coincide with a climatological disaster (drought or heat wave) in the same country even without applying a disaster damage threshold (7% for the EF country grouping). For hydrological events, only those with strongest impact yield statistically significant results, albeit at a low precursor coincidence rate. Also, we only find significant precursor coincidence for meteorological disasters for EF with a low coincidence rate.

The same analysis has been performed for the trigger test quantifying to what degree armed-conflict outbreaks coincide with or follow disasters (Fig. 4). Again, we find the signal for coincidences within the same month and climatological disasters to be most robust with the largest statistically significant coincidence rate for the EF country group. However, trigger coincidences have only been identified for 2.5% of all climatological events and about 2% of all disasters above a 1% relative GDP threshold for the EF country grouping and are not robust at the global scale.

Discussion

The question whether or not climate-related factors have significantly contributed to recent armed-conflict outbreaks has been heavily disputed in the scientific literature (33–38). Although a sequence of studies has suggested that a large number of outbreaks of armed conflicts in modern as well as premodern times have been associated with climatic variability (33, 36, 37, 39–41), the robustness of these findings and underlying mechanisms are controversially discussed (10, 37, 42, 43). Other literature that assessed the influence of climate signals on armed-conflict outbreak risk did not report a robust connection (9, 44, 45).

A clear shortcoming of most studies investigating the relation between climate change and armed conflicts is that they focus solely on meteorological indices such as temperature or precipitation time series (9, 39–42, 46), thereby neglecting the crucial importance of vulnerability and exposure for the impacts of climate hazards (35, 47). This might be one reason for the substantial disagreement on the matter in the literature. Moving beyond purely meteorological indices toward the development of composite indices accounting for vulnerability and exposure to climate change, as well as conflict risk provides a promising way forward to reconcile this debate (48, 49).

Our ECA approach, based on disaster occurrence characterized by the economic impact of a climate-related event instead of a meteorological index, accounts to some extent for the effects of vulnerability and exposure. However, some potential caveats need to be considered. Economic losses as measured relative to GDP are of limited relevance in assessing disaster impacts in the most vulnerable countries, as disaster-related losses are difficult to quantify and loss of lives and livelihoods may substantially outweigh economic losses. At the same time, damages by disasters that are not directly affecting economic assets but rather living conditions and subsistence agriculture, such as droughts, are difficult to quantify in economic terms (32). These shortcomings of the economic indicators may explain why we find robust significant relationships down to low damage threshold levels as well as the apparent insensitivity to the threshold level for climatological events (compare Fig. 3). A second shortcoming is associated with the country-level resolution of our study that can impede the assessment of potential relations

Reproduced from: C. F. Schleussner et al., Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries, Proc. Natl. Acad. Sci. U. S. A., vol. 113, no. 33, pp. 9216-9221, 2016, doi: 10.1073/pnas.1601611113.

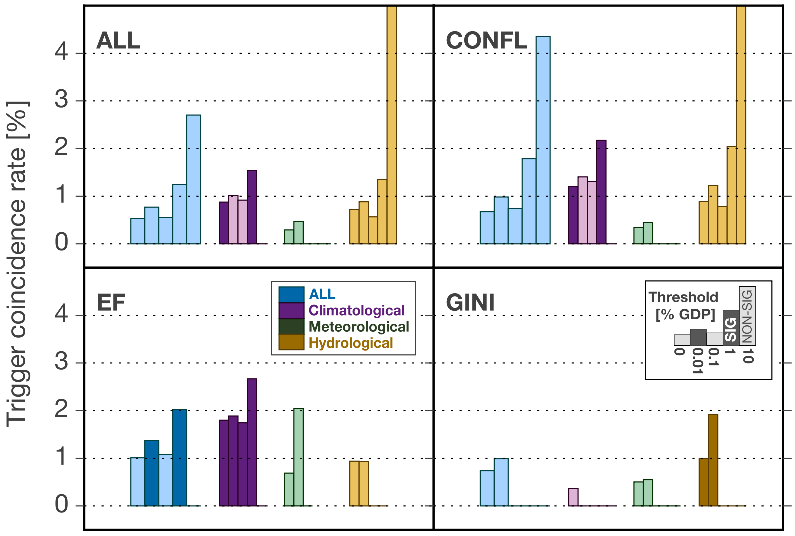


Fig. 4. Results of ECA for the trigger test based on the occurrence of disasters that coincide with an armed-conflict outbreak within the same month (*Materials and Methods*). Results for four different country groupings are resolved in four individual panels, whereas results for different disaster types are indicated by the color coding. Coincidence rates are displayed for different damage threshold levels by individual bars with increasing damage threshold from left to right. For some threshold levels, the trigger coincidence rate is zero. Filled segments indicate coincidence rates that are significant at the 95% level. Note that the coincidence rates are one order of magnitude smaller than for the risk enhancement test depicted in Fig. 3.

between disasters and armed-conflict outbreaks happening in different parts of the country. Although a higher resolution would indeed improve our analysis, the current country aggregation does not undermine the validity of our test as a larger number of possibly disconnected disasters and armed-conflict outbreaks on a country level will only lead to higher significance thresholds (*Materials and Methods*) and, consequently, to a more conservative test.

The results for the different country groupings depend not only on the ad hoc selection of group sizes (although our findings for EF are largely robust for different group sizes; compare *SI Appendix*, Fig. S1) but also on the index chosen. Although widely used, the classification following general indicators such as GINI or EF has led to inconsistent results in the conflict literature and it has been shown that theoretically informed country profiles combining multiple factors and relating them to dimensions of power sharing are much better predictors of armed-conflict outbreaks (10). Specifically, discriminatory political and power-sharing systems along ethnic boundaries have been found to be of key relevance (10). Thereby, a refinement of our analysis based on indices reflecting ethnic inclusiveness in power sharing might be promising for further research.

Commonly, high ethnic separation on a country level coincides with other potential sources of conflict such as economic inequality or poverty, which makes it difficult to disentangle their specific effects (50). However, the results for alternative groupings (e.g., for inequality, poverty, and conflict proneness) are much less robust than those for EF, despite a substantial overlap in the actual country groupings (compare Fig. 2). Although the country grouping approach as applied here does not allow for a direct quantification of the driver's importance, our results imply that the mechanisms specific to EF and conflict outbreak discussed above may play a significant role for armed-conflict outbreak following a natural disaster (18). Thereby, it is not the domain-specific factors, EF, or natural disasters occurrence alone, but their interplay that results in enhanced risk of armed-conflict outbreak. Besides our robust findings of risk enhancement, we report no further indications that natural disasters are causing armed-conflict outbreaks in a more direct manner (based on the trigger test). Thereby, our results do not support attempts of single-factor attribution of conflict outbreaks to disaster occurrence.

Unlike development-related factors such as poverty and inequality, ethnic fractionalization of societies cannot be overcome via economic development alone. As a consequence, country-specific risks may prevail over the next decades independently of the countries' state of development, if no robust progress in ethnic inclusiveness regarding power sharing is achieved (10). Among the most fractionalized countries are many African as well as Central Asian nations (compare Fig. 2), which makes these regions potential hot spots of armed-conflict outbreak risk enhancement due to climate-related natural disasters. Climate projections indicate a substantial increase in extreme event hazards in these regions and most of the affected countries are also characterized by high vulnerability and low adaptive capacity, which renders them particularly susceptible to high-impact climate-related natural disasters (51, 52). Projections of overall conflict risk up to 2050 based on a multifactorial analysis also find these regions to be particularly endangered (12), which highlights the relevance of our findings in the wider context of conflict prevention and development.

The robust finding of armed-conflict outbreak risk enhancement for climatological events globally points towards increased risks due to a projected drying trend in already drought-prone regions such as Northern Africa and the Levant (53). Recent analyses of the societal consequences of droughts in Syria and Somalia indicate that such climatological events may have already contributed to armed-conflict outbreaks or sustained conflicts in both countries (54–57). Similarly, a prolonged drought might have contributed negatively to the ongoing conflicts in Afghanistan (58). Further destabilization of Northern Africa and the Levant may have widespread effects by triggering migration flows to neighboring countries and remote migrant destinations such as the European Union. Although not highly ethnically fractionalized following the ad hoc threshold classification applied here, ethnic identities also appear to play a prominent role in the ongoing civil wars in Syria and Iraq (18). It is clear that the roots of these conflicts, as for armed conflicts in general, are case specific and not directly associated with climate-related natural disasters. Nevertheless, such disruptive events have the potential to amplify already existing societal tensions and stressors and thus to further destabilize several of the world's most conflict-prone regions (12, 31).

Reproduced from: C. F. Schleussner et al., Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries, Proc. Natl. Acad. Sci. U. S. A., vol. 113, no. 33, pp. 9216-9221, 2016, doi: 10.1073/pnas.1601611113.

Materials and Methods

Data Sources.

Natural disaster database. The analysis of disaster damages is based on the NatCatSERVICE database from Munich Re (32) developed for the private sector, which is available upon request from the Munich Re NatCatSERVICE. This database provides state-of-the-art estimates of economic damages connected to natural hazards. The database comprises the 1980–2010 period and gives estimates for total economic damages based on internal estimates and third-party sources. It contains about 18,000 climate-related events for that period. Damage events are classified according to the nature of the underlying natural hazard and include also geophysical events such as earthquakes, which are excluded from our analysis (see *SI Appendix*, Table S1 for an overview on the climate-related natural hazards and their classification). To account for country-specific economic conditions, the absolute damages are considered relative to the countries' annual GDP (International Monetary Fund database; https://www.imf.org/external/data.htm), which allows for the analysis of climate-related natural hazards dependent on their destructiveness in economic terms. All damages are deflated to 2010 US dollars.

Armed-conflict database. Data on armed conflicts are taken from the openly available UCDP/PRIO Armed Conflict Dataset (30, 31) (www.pcr.uu.se/research/ucdp/datasets/ucdp_prio_armed_conflict_dataset/). This dataset counts all incidences with more than 25 battle-related deaths globally, both interstate and intrastate conflicts. Conflict outbreaks are counted on a yearly basis, for each dyad of conflicting parties (either interstate or intrastate). For ongoing conflicts, each new outbreak is included when preceded by at least 24 months of nonconflict. Interstate conflicts are treated separately and coincidences are counted if at least one of the countries has been hit by a disaster within the coincidence interval and above the damage threshold. Conflicts involving multiple countries (such as US-led coalitions in Afghanistan and Iraq in the 2000s) are excluded. The dataset includes 241 conflict outbreaks over the 1980–2010 period for both interstate and intrastate conflicts globally.

Country classification. The country classification in terms of ethnic as well as religious fractionalization is based on indices developed by Alesina et al. (29) and the Gini coefficient is based on World Bank data and averaged over the 1980–2010 period (World Bank database; data.worldbank.org/indicator/). For both indices, the 50 countries with the highest values are used. For further country classifications, see *SI Appendix*, Table S3.

Method Description: ECA. ECA is a method tailored for quantifying and testing statistical interrelationships between event series while allowing to specify explicitly the coincidence interval, lag, and directionality (in terms of precursor and trigger coincidences) of these interrelationships (24). In this study, we perform two coincidence tests: (*i*) the risk enhancement test, which is based on armed-conflict outbreak and tests for coincidences of natural disasters co-occurring with or preceding conflict events, and (*ii*) the trigger test based on climate-related natural disaster occurrence, which tests for coincidences with armed-conflict outbreaks following or co-occurring with a disaster event (24). Both tests differ with regard to the considered set of countries and the definition of the coincidence interval, but otherwise the same methodology is applied. We analyze countrywise coincidences between armed-conflict outbreaks at times $t_i^{c,k}$ $(i = 1, \ldots, N_{c,k})$ and disaster events at times $t_j^{d,k}(\varepsilon)$ $(j = 1, \ldots, N_{d,k}(\varepsilon))$ within a coincidence interval ΔT (Fig. 1) for a group of countries G, where $k \in G$ is a country index. $N_{c,k}$ and $N_{d,k}(\varepsilon)$ denote the numbers of armed conflicts and disasters for a given country k, respectively. The disaster events are filtered by a damage threshold ε measured in units relative to annual GDP.

The risk enhancement test is based on the aggregated precursor coincidence rate $r_p^G(\Delta T, \varepsilon)$ (24) measuring the fraction of conflicts in country group G that were preceded by at least one disaster of the strength of at least ε in the same country and that occurred at most at time ΔT before the conflict started:

$$r_p^G(\Delta T, \varepsilon) = \frac{\sum_{k\in G}\sum_{i=1}^{N_{c,k}} \Theta\left[\sum_{j=1}^{N_{d,k}(\varepsilon)} 1_{[0,\,\Delta T]}\left(t_i^{c,k} - t_j^{d,k}(\varepsilon)\right)\right]}{\sum_{k\in G} N_{c,k}}, \qquad [1]$$

where $\Theta(\cdot)$ is the Heaviside function [here defined as $\Theta(x) = 0$ for $x \le 0$ and $\Theta(x) = 1$ otherwise] and $1_I(\cdot)$, the indicator function of the interval I [defined as $1_I(x) = 1$ for $x \in I$ and $1_I(x) = 0$ otherwise]. Note that, according to this definition, multiple disasters preceding a given conflict within the coincidence interval are counted only once. In turn, the trigger test is based on computing aggregated trigger coincidence rates (24):

$$r_t^G(\Delta T, \varepsilon) = \frac{\sum_{k\in G}\sum_{j=1}^{N_{d,k}(\varepsilon)} \Theta\left[\sum_{i=1}^{N_{c,k}} 1_{[0,\,\Delta T]}\left(t_i^{c,k} - t_j^{d,k}(\varepsilon)\right)\right]}{\sum_{k\in G} N_{d,k}(\varepsilon)}, \qquad [2]$$

measuring the fraction of disasters of a strength of at least ε in country group G that were followed by at least one conflict that occurred at most a time ΔT after the disaster onset in the same country.

The temporal resolution of the analysis is limited to monthly values, which accounts for both dating uncertainties in the conflict database as well as in disaster onsets (as in, e.g., droughts). For temporally extended disaster events, the start date is used. Although certain events such as heat waves and in particular droughts can last for several months, an analysis using the end dates of such temporally extended disasters (not shown) does not exhibit significant coincidence rates. To assess the statistical robustness of our findings, independent Poisson processes are assumed for both the disaster as well as the conflict outbreak event series at the individual country level, conserving the event rates $N_{c,k}/T$ and $N_{d,k}(\varepsilon)/T$, respectively (23). Here, T denotes the total time span covered by both event series. The corresponding null hypothesis (NH) to be tested is that the observed coincidence rates for a group of countries G occur due to chance alone. To perform this test, Monte Carlo simulation is applied for generating M pairs of surrogate event series. Event rates for each country $k \in G$ are conserved by uniformly and independently drawing $N_{c,k}, N_{d,k}(\varepsilon)$ event timings from the considered period 1980–2010 to compute a test distribution of coincidence rates $p(r^G)$ using Eqs. 1 and 2. For each considered country grouping, $M = 1{,}000$ ensemble members are generated and a 95% significance level is applied for the rejection of the NH of coincidence rates arising due to chance alone. No significance assessments are made, if the absolute number of coincidences counted is smaller than 2.

A variety of approaches related to ECA is applied in the neurosciences for investigating statistical interrelationships between neuronal spike trains (26). Among others, event synchronization (25) has been widely used for studying climatological extreme events in various contexts (60, 61). Donges et al. (24) provide a more detailed discussion of ECA in comparison with related approaches.

It should be noted that the statistically significant coincidence rates observed in this study could in principle be due to a hidden common cause that affects the timing of both climate-related disasters and armed-conflict outbreaks. Although the existence of such a root cause cannot be ruled out a priori, there is no obvious hypothesis available on what a hidden common cause or common driver could be in the setting of our study. If event or other data on candidate processes is available, extensions of ECA such as conditional ECA could be applied to study common driver effects (62). Alternatively, recurrence-based methods proposed for discovering hidden common causes in the case of bivariate standard time series (63) could be adapted for event time series in future research.

The software (Python scripts) and openly available data used for performing the analysis presented in this paper have been made available at www.pik-potsdam.de/research/publications/pnas/Schleussner_et_al_2016_PNAS_scripts.zip.

ACKNOWLEDGMENTS. We thank Wolfgang Lucht for fruitful discussions and his continuous support. Jobst Heitzig is acknowledged for valuable comments and his contributions to the development of the method of ECA. This work was conducted in the framework of Potsdam Institute for Climate Impact Research's flagship project on Coevolutionary Pathways in the Earth System (COPAN). Munich Re is acknowledged for providing access to their NatCatSERVICE database. We appreciate funding by the Humboldt University Berlin (Integrative Research Institute on Transformations of Human–Environment Systems Fellowship), the Stordalen Foundation (via the Planetary Boundary Research Network PB.net), the Earth League's EarthDoc Programme, by the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety (11_II_093_Global_A_SIDS and LDCs), and the German Federal Ministry for Education and Research [Project GLUES and Young Investigators Group CoSy-CC^2 (Grant 01LN1306A)].

1. Tainter J (1990) *The Collapse of Complex Societies.* New Studies in Archeology (Cambridge Univ Press, Cambridge, UK).
2. Büntgen U, et al. (2011) 2500 years of European climate variability and human susceptibility. *Science* 331(6017):578–582.
3. Cullen HM, et al. (2000) Climate change and the collapse of the Akkadian empire: Evidence from the deep sea. *Geology* 28(4):379–382.
4. deMenocal PB (2001) Cultural responses to climate change during the late Holocene. *Science* 292(5517):667–673.
5. Drysdale R, et al. (2006) Late Holocene drought responsible for the collapse of Old World civilizations is recorded in an Italian cave flowstone. *Geology* 34(2):101–104.
6. Haug GH, et al. (2003) Climate and the collapse of Maya civilization. *Science* 299(5613):1731–1735.
7. Kennett DJ, et al. (2012) Development and disintegration of Maya political systems in response to climate change. *Science* 338(6108):788–791.
8. Donges JF, et al. (2015) Nonlinear regime shifts in Holocene Asian monsoon variability: Potential impacts on cultural change and migratory patterns. *Clim Past* 11:709–741.

Reproduced from: C. F. Schleussner et al., Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries, Proc. Natl. Acad. Sci. U. S. A., vol. 113, no. 33, pp. 9216-9221, 2016, doi: 10.1073/pnas.1601611113.

9. Slettebak RT (2012) Don't blame the weather! Climate-related natural disasters and civil conflict. *J Peace Res* 49(1):163–176.
10. Buhaug H, Cederman LE, Gleditsch KS (2014) Square pegs in round holes: Inequalities, grievances, and civil war. *Int Stud Q* 58:418–431.
11. Fearon JD (2010) *Governance and Civil War Onset. Background Paper, World Development Report 2011* (World Bank, Washington, DC).
12. Hegre H, Karlsen J, Nygård HM, Strand H, Urdal H (2013) Predicting armed conflict, 2010–2050. *Int Stud Q* 57(2):250–270.
13. Collier P, Hoeffler A (2004) Greed and grievance in civil war. *Oxf Econ Pap* 56:563–595.
14. Collier P, Hoeffler A, Rohner D (2009) Beyond greed and grievance: Feasibility and civil war. *Oxf Econ Pap* 61(1):1–27.
15. Duffy Toft M (2002) Indivisible territory, geographic concentration, and ethnic war. *Secur Stud* 12(2):82–119.
16. Mishali-Ram M (2006) Ethnic diversity, issues, and international crisis dynamics, 1918–2002. *J Peace Res* 43(5):583–600.
17. Wegenast TC, Basedau M (2013) Ethnic fractionalization, natural resources and armed conflict. *Confl Manage Peace Sci* 31(4):432–457.
18. Denny EK, Walter BF (2014) Ethnicity and civil war. *J Peace Res* 51(2):199–212.
19. Frank R, Rainer I (2012) Does the leader's ethnicity matter? Ethnic favoritism, education, and health in sub-Saharan Africa. *Am Polit Sci Rev* 106:294–325.
20. Eifert B, Miguel E, Posner DN (2010) Political competition and ethnic identification in Africa. *Am J Pol Sci* 54(2):494–510.
21. Anthias F (2007) Ethnic ties: Social capital and the question of mobilisability. *Sociol Rev* 55(4):788–805.
22. Olsson L, et al. (2014) Livelihoods and poverty. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel of Climate Change*, eds Field CB, et al. (Cambridge Univy Press, Cambridge, UK), pp 793–832.
23. Donges JF, et al. (2011) Nonlinear detection of paleoclimate-variability transitions possibly related to human evolution. *Proc Natl Acad Sci USA* 108(51):20422–20427.
24. Donges JF, Schleussner CF, Siegmund JF, Donner RV (2016) Event coincidence analysis for quantifying statistical interrelationships between event time series: On the role of flood events as triggers of epidemic outbreaks. *Eur Phys J Spec Top* 225:469–485.
25. Quian Quiroga R, Kreuz T, Grassberger P (2002) Event synchronization: A simple and fast method to measure synchronicity and time delay patterns. *Phys Rev E Stat Nonlin Soft Matter Phys* 66(4 Pt 1):041904.
26. Brown EN, Kass RE, Mitra PP (2004) Multiple neural spike train data analysis: State-of-the-art and future challenges. *Nat Neurosci* 7(5):456–461.
27. Rammig A, et al. (2015) Coincidences of climate extremes and anomalous vegetation responses: Comparing tree ring patterns to simulated productivity. *Biogeosciences* 12(2):373–385.
28. Siegmund JF, Wiedermann M, Donges JF, Donner RV (2015) Impact of climate extremes on wildlife plant flowering over Germany. *Biogeosciences Discuss* 12:18389–18423.
29. Alesina A, Devleeschauwer A, Easterly W, Kurlat S, Wacziarg R (2003) Fractionalization. *J Econ Growth* 8:155–194.
30. Gleditsch NP, Wallensteen P, Eriksson M, Sollenberg M, Strand H (2002) Armed conflict 1946–2001: A new dataset. *J Peace Res* 39(5):615–637.
31. Themnér L, Wallensteen P (2012) Armed conflicts, 1946–2011. *J Peace Res* 49(4):565–575.
32. Wirtz A, Kron W, Löw P, Steuer M (2014) The need for data: Natural disasters and the challenges of database management. *Nat Hazards* 70(1):135–157.
33. Burke MB, Miguel E, Satyanath S, Dykema JA, Lobell DB (2009) Warming increases the risk of civil war in Africa. *Proc Natl Acad Sci USA* 106(49):20670–20674.
34. Buhaug H (2010) Climate not to blame for African civil wars. *Proc Natl Acad Sci USA* 107(38):16477–16482.
35. Scheffran J, Brzoska M, Kominek J, Link PM, Schilling J (2012) Climate change and violent conflict. *Science* 336(6083):869–871.
36. Hsiang SM, Burke M (2014) Climate, conflict, and social stability: What does the evidence say? *Clim Change* 123(1):39–55.
37. Hsiang SM, Meng KC (2014) Reconciling disagreement over climate-conflict results in Africa. *Proc Natl Acad Sci USA* 111(6):2100–2103.
38. Buhaug AH, et al. (2014) One effect to rule them all? A comment on climate and conflict. *Clim Change* 127(3-4):391–397.
39. Cane MA, et al. (2014) Temperature and violence. *Nat Clim Chang* 4(4):234–235.
40. Hsiang SM, Meng KC, Cane MA (2011) Civil conflicts are associated with the global climate. *Nature* 476(7361):438–441.
41. Hsiang SM, Burke M, Miguel E (2013) Quantifying the influence of climate on human conflict. *Science* 341(6151):1235367.
42. Gleditsch NP (2012) Whither the weather? Climate change and conflict. *J Peace Res* 49(1):3–9.
43. O'Loughlin J, Linke AM, Witmer FDW (2014) Modeling and data choices sway conclusions about climate-conflict links. *Proc Natl Acad Sci USA* 111(6):2054–2055.
44. Nel P, Righarts M (2008) Natural disasters and the risk of violent civil conflict. *Int Stud Q* 52(1):159–185.
45. Bergholt D, Lujala P (2012) Climate-related natural disasters, economic growth, and armed civil conflict. *J Peace Res* 49(1):147–162.
46. Theisen OM, Gleditsch NP, Buhaug H (2013) Is climate change a driver of armed conflict? *Clim Change* 117(3):613–625.
47. IPCC (2014) Summary for policy makers. *Climate Change 2014: Impacts, Adaptation and Vulnerability. Contributions of the Working Group II to the Fifth Assessment Report*, eds Field CB, et al. (Cambridge Univ Press, Cambridge, UK), pp 1–32.
48. Ide T, et al. (2014) On exposure, vulnerability and violence: Spatial distribution of risk factors for climate change and violent conflict across Kenya and Uganda. *Polit Geogr* 43:68–81.
49. Scheffran J, Brzoska M, Kominek J, Link PM, Schilling J (2012) Disentangling the climate-conflict nexus: Empirical and theoretical assessment of vulnerabilities and pathways. *Rev Eur Stud* 4(5):1–13.
50. Fearon JD, Laitin DD (2003) Ethnicity, insurgency, and civil war. *Am Polit Sci Rev* 97(1):75–90.
51. Niang I, et al. (2014) Africa. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel of Climate Change*, eds Barros VR, et al. (Cambridge Univ Press, Cambridge, UK), pp 1199–1265.
52. Hijioka Y, et al. (2014) Asia. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel of Climate Change*, eds Barros VR, et al. (Cambridge Univ Press, Cambridge, UK), pp 1327–1370.
53. Field C, et al., eds (2012) *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation* (Cambridge Univ Press, Cambridge, UK).
54. Kelley CP, Mohtadi S, Cane MA, Seager R, Kushnir Y (2015) Climate change in the Fertile Crescent and implications of the recent Syrian drought. *Proc Natl Acad Sci USA* 112(11):3241–3246.
55. Werrell CE, Femia F, Sternberg T (2015) Did we see it coming? State fragility, climate vulnerability, and the uprisings in Syria and Egypt. *SAIS Rev* 35(1):29–46.
56. Gleick PH (2014) Water, drought, climate change, and conflict in Syria. *Weather Clim Soc* 6(3):331–340.
57. Maystadt JF, Ecker O (2014) Extreme weather and civil war: Does drought fuel conflict in Somalia through livestock price shocks? *Am J Agric Econ* 96:1157–1182.
58. Parenti C (2015) Flower of war: An environmental history of opium poppy in Afghanistan. *SAIS Rev* 35(1):183–200.
59. Theisen OM, Holtermann H, Buhaug H (2011) Climate wars? Assessing the claim that drought breeds conflict. *Int Secur* 36(3):79–106.
60. Malik N, Bookhagen B, Marwan N, Kurths J (2012) Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks. *Clim Dyn* 39(3-4):971–987.
61. Boers N, Bookhagen B, Marwan N, Kurths J, Marengo J (2013) Complex networks identify spatial patterns of extreme rainfall events of the South American monsoon system. *Geophys Res Lett* 40(16):4386–4392.
62. Siegmund JF, et al. (2016) Meteorological drivers of extremes in daily beech, oak and pine stem diameter variations in northeastern Germany: An event coincidence analysis. *Front Plant Sci* 7:00733.
63. Hirata Y, Aihara K (2010) Identifying hidden common causes from bivariate time series: A method using recurrence plots. *Phys Rev E Stat Nonlin Soft Matter Phys* 81(1 Pt 2):016203.

Reproduced from: C. F. Schleussner et al., Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries, Proc. Natl. Acad. Sci. U. S. A., vol. 113, no. 33, pp. 9216-9221, 2016, doi: 10.1073/pnas.1601611113.

EPJ manuscript No.
(will be inserted by the editor)

Dose-response function approach for detecting spreading processes in temporal network data

Exploring social contagion in the Copenhagen Networks Study

Jonathan F. Donges[1,2,*,a], Jakob Lochner[1,3,*], Jobst Heitzig[1], Niklas Kitzmann[1,4], Sune Lehmann[5,6], Marc Wiedermann[1], and Jürgen Vollmer[3]

[1] Earth System Analysis & Complexity Science, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, 14473 Potsdam, Germany, EU
[2] Stockholm Resilience Centre, Stockholm University, 10691 Stockholm, Sweden, EU
[3] Institute for Theoretical Physics, University of Leipzig, 04103 Leipzig, Germany, EU
[4] Institute for Physics and Astronomy, University of Potsdam, 14476 Potsdam, Germany, EU
[5] Department of Applied Mathematics and Computer Science, Technical University of Denmark, Lyngby, Denmark, EU
[6] Department of Sociology, University of Copenhagen, Copenhagen, Denmark, EU
* The first two authors share the lead authorship.

Abstract. Spreading or complex contagion processes on networks are an important mechanistic foundation of tipping dynamics and other nonlinear phenomena in complex social, ecological and technological systems. Increasing amounts of temporal network data are now becoming available to study such spreading processes of behaviours, opinions, ideas, diseases, innovations or technologies and to test hypotheses regarding their specific properties. To this end, we here present a methodology based on dose-response functions and hypothesis testing using surrogate data sets. We demonstrate this methodology for synthetic temporal network data generated by the adaptive voter model. Furthermore, we apply it to empirical temporal network data from the Copenhagen Networks Study. This data set provides a physically-close-contact network between university students participating in the study over the course of three months. We study the potential spreading dynamics of the health-related behaviour "regularly going to the fitness studio" on this network. Based on a hierarchy of surrogate data models, we find that the empirical data neither provide significant evidence for an influence of a dose-response-type network spreading process, nor significant evidence for homophily. The empirical dynamics in exercise behaviour are likely better described by individual features such as the disposition towards the behaviour, and the persistence to maintain it, as well as external influences affecting the whole group, and the non-trivial network structure. The proposed methodology is generic and promising also for applications to other data sets and traits of interest.

1 Introduction

Spreading and contagion processes shape the dynamics of diverse complex ecological, societal and technological systems studied in many fields of research [1–3]. Examples include biological infections [4, 5] such as the spreading of the COVID-19 pandemic [6], cascading failures in interdependent infrastructure systems [7], diffusion of innovations and technologies [8–10], social norms [11] and other social, political and technological innovations relevant for sustainability transition and rapid decarbonisation [12–15], political changes [16], or religious missionary work [17, 18]. These spreading processes on complex networks often give rise to nonlinear dynamics and the emergence of macroscopic phenomena, such as phase transitions and tipping points that separate qualitatively different dynamical regimes [19]; for example, a transition between regimes where a local infection or innovation is locally contained, and those where it spreads globally to a large part of the network [1, 2, 10, 20, 21]. Furthermore, spreading processes can interact with the underlying complex network structures, e.g. through the process of homophily, giving rise to complex coevolutionary feedbacks between dynamics on and structure of these networks [22–25]. Better understanding of such complex spreading processes, based on improved methods for data analysis and modelling, is highly relevant for finding robust approaches to influence, manage, govern or control their dynamics. This way, harmful impacts may be avoided, or desirable outcomes reached, e.g. for containing pandemic outbreaks [6, 26, 27], preventing cascading failures in power grids [7, 28], or fostering the spreading of social-cultural-technological innovations towards a rapid sustainability transformation [12–14, 19].

[a] e-mail: donges@pik-potsdam.de

In recent years, temporal network data has become more abundantly available from social media platforms such as Facebook and Twitter, or long-term health studies such as the Framingham Heart Study that have been leveraged for studying spreading and contagion processes, e.g. in the dynamics of obesity [29], smoking [30], happiness [31], loneliness [32], alcohol consumption [33], depression [34], divorce [35], emotional contagion [36] and political mobilisation [37]. So far such studies of empirical temporal network data mainly relied on standard statistical methods such as generalised linear models, generalised estimating equations or spatial autoregressive models. However, these methods are typically not well-equipped to deal with network dependencies [38]. Furthermore, analogous to the problem of identifying causal associations in multivariate time series data [39, 40], there are challenges in extracting possible causal effects induced by contagion processes, and in separating their imprints from other mechanisms such as homophilic rewiring of network structure, common external forcing from the system's environment and other confounding effects. After all, most studies rely on observational data and not on controlled experiments [38].

Here, we contribute to this field by developing a methodology for the analysis of complex spreading processes in temporal network data sets based on dose response functions (DRFs) that have been used in the theoretical description of simple and complex contagion processes [2, 20]. Among others, they have been applied to the study of behavioural contagion in animal systems such as startling cascades in fish schools [41] and the spread of information on social media networks [42]. Dose response functions encode a network nodes' probability of being infected with a new trait, given the level of exposure to this trait in its network neighbourhood. We propose an algorithm including Gaussian filtering to robustly estimate DRFs from synthetic and empirical temporal network data, including the possibility of propagating various types of uncertainties. In order to test for the possibility of an actual causal spreading process being involved in generating the data, and to identify confounding effects, we also develop a hierarchy of temporal network surrogate models. They enable us to investigate which features and structures in the data are possibly sufficient to explain the obtained dose response functions.

We apply this methodology to synthetic data from the adaptive voter model as a proof-of-concept, and to empirical observational temporal network data from the Copenhagen Networks Study. Based on the latter we analyse the spreading dynamics of the illustrative behaviour of "regularly going to the fitness studio" on a physically-close-contact network between university students participating in the study over the course of three months. We do not find robust evidence of a causal spreading process underlying the observed dynamics. This suggests that possible social contagion effects in this context are very limited, and dominated by other factors or shadowed by excessive noise. This is in agreement with findings from health behaviour psychology [43]. Hence, this first application study suggests that the proposed methodology is generic and promising for investigations of other data sets and possibly spreading traits of interest.

This paper is structured as follows: we first introduce the synthetic and empirical temporal network data sets, obtained from the adaptive voter model and the Copenhagen Network Study, respectively (Sect. 2). In a next step, we describe the methodology developed here for data analysis, including estimating dose response functions and generating surrogate data sets for testing hypotheses on underlying data generating processes (Sect. 3). Finally, we report results obtained for the synthetic and empirical data sets (Sect. 4), discuss these findings and conclude (Sect. 5).

2 Data

Here we describe the data sets used in this study to test our proposed dose-response function methodology. The data has the form of temporal networks (Sect. 2.1), it includes synthetic temporal network data generated by the adaptive voter model (Sect. 2.2) and empirical temporal network data from the Copenhagen Networks Study (Sect. 2.3).

2.1 Temporal social networks

The data sets investigated in this work are structured as temporal networks $\mathcal{G}(t)$ with a fixed number of nodes N and a time-dependent set of links described by the adjacency matrix $A_{ij}(t)$, where $i, j \in \{1, \dots, N\}$ [44], sampled at discrete time steps t. In addition, node traits $o_i(t)$ are time-dependent as well, for example encoding changing opinions or behaviours.

2.2 Synthetic temporal network data: adaptive voter model

One prototypical model of temporal network dynamics is the adaptive voter model (AVM) [22] that incorporates core processes in social systems, i.e., homophily [45] and social learning of traits [46]. As such, the AVM can be interpreted as a straightforward generalisation of the so-called voter model [47] to any prescribed initial social network topology and the ability of the represented individuals to deliberately change their neighbourhood structure. It thereby aims to explain the emergence of like-minded communities within a larger social network and the extent to which individuals (i) become like-minded because of shared social ties or (ii) form such social ties because they are like-minded.

Specifically, the model considers a temporal network $\mathcal{G}(t)$ with a fixed number of N nodes and M links. Each node v_i holds one of Γ opinions or traits o_i that are initially distributed at random among them. The M links are initially distributed uniformly at random as well, thus mimicking the configuration of an Erdős–Rényi graph. At each

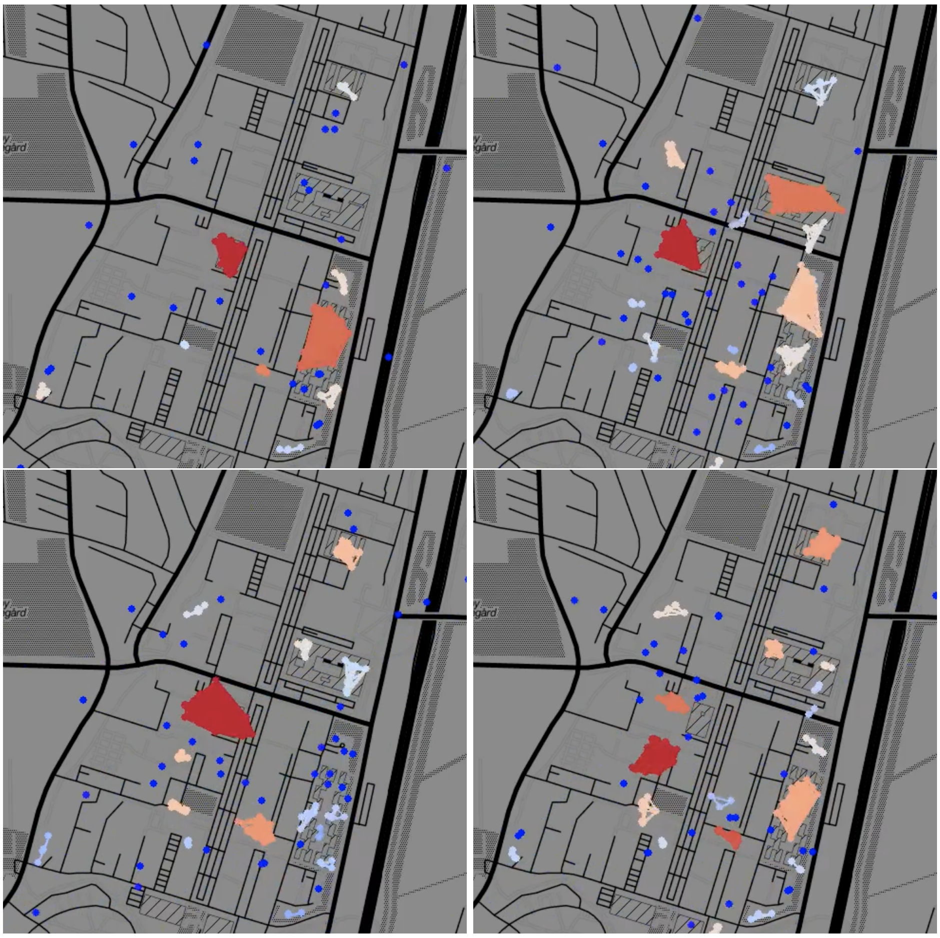


Fig. 1: Temporal network snapshots throughout a typical day during the first semester of the Copenhagen Networks Study. Each dot represents an individual, colour coded according to cluster size from single nodes (dark blue) to large clusters (dark red). Node clusters evident in the snapshots correspond to students engaging in joint activities, such as lectures or eating lunch in a cafeteria.

discrete time step t, a single node v_i with opinion or trait o_i is randomly chosen. If its degree k_i, i.e. the number of directly connected neighbours, is non-zero, either of two processes takes place:

1. *Homophilic rewiring*. With fixed probability φ we select one of the edges that are attached to v_i and move its other end to a randomly selected node v_k that holds the same trait o_k as v_i, and is not connected to v_i yet. v_i thereby *adapts* its neighbourhood structure to align more with its own trait o_i.
2. *Social learning*: Otherwise, with fixed probability $1 - \varphi$ we pick a random neighbour v_j of v_i and set v_i's trait equal to that of v_j, i.e., $v_i \leftarrow v_j$. Hence, v_i *imitates* the trait o_k of v_k to become more alike to its immediate neighbourhood.

The model reaches a steady state once only one trait per connected network component remains. In this case, no additional updates to the nodes' states or their neighbourhood structure are possible. The fixed probability φ is a model parameter that allows to scale the relative frequencies of imitation and adaptation events. For $\varphi = 0$ only imitation, and for $\varphi = 1$ only adaptation takes place. The model displays a phase transition at intermediate values of φ where the system's steady state qualitatively shifts from a large connected component of a single remaining trait to a fictionalised configuration of multiple disconnected components that each show distinct predominant traits [22].

In our specific study we set the number of nodes to $N = 850$, the number of edges to $M = 5724$ and the number of traits to $\Gamma = 2$ to ensure consistency with the empirical data from the Copenhagen Networks Study (CNS), see below.

2.3 Empirical temporal network data: Copenhagen Networks Study

In the following, we present the Copenhagen Networks Study as our main empirical data source (Sect. 2.3.1) and describe the methodology used for extracting a temporal social network with time-dependent node traits from this data set (Sect. 2.3.2).

2.3.1 Description of data sources

The data analysed here originates from the Copenhagen Networks Study (CNS) [48, 49]. CNS was carried out from 2012–2016 and focused on collecting temporal network and demographic data on a densely interconnected cohort of nearly 1000 individuals. In order to collect the temporal network information, the study handed out state-of-the-art smartphones to consenting freshman students at the Technical University of Denmark. Specifically the study collected information on networks of physical proximity (using Bluetooth signals), phone calls, text messages, and online social networks. In addition to the network data, the study also collected information on the participants' mobility, using the phones' GPS sensors – and demographic and personality data, using questionnaires. The study was approved by the Danish Data Protection agency, the appropriate legal entity in Denmark. In terms of research, data from CNS have been used in a number of contexts e.g. epidemiology [50–52], mobility research [53, 54], network science [55, 56], studies of gender-related behaviour [57], and education research [58, 59].

In addition to the data from the Copenhagen Networks Study, and in view of our aim to investigate the illustrative behaviour "regularly going to the fitness studio", a data set was generated with the locations of fitness studios in the vicinity of Copenhagen. The studios were selected from the locations provided by Open Street Map [60] and listed with the keys 'leisure=fitness_center' or 'sport=fitness'. A comprehensive list of all considered studios can be found in Appendix A.

2.3.2 Generation of empirical temporal social network

The empirical temporal social network is generated as a physically-close-contact network between the study's participants. A network edge is created when two participants are in close proximity to each other at a time t. The network's adjacency matrix $A_{ij}(t)$ is then defined as

$$A_{ij}(t) = \begin{cases} 1\ , & |s_{ij}(t)| > 80\,\text{dBm} \\ 0\ , & \text{otherwise} \end{cases} , \tag{1}$$

where time t is in units of days and $s_{ij}(t)$ is the maximum Bluetooth signal strength between participants i and j measured during day t. The threshold $80\,$dBm corresponds to a distance of about $2\,$m and maximises the ratio of social interactions to transient and unimportant connections [61].

In order to minimise noise from the beginning and end periods of data collection, in this study we focus on the period from the first of February 2014 to the end of April 2014, which corresponds to the spring semester and is in the middle of the "SensibleDTU 2013" data collection, the second deployment of CNS. Furthermore, a minimum level of social interaction is essential for our study. Therefore, participants who had no or very few contact events were removed from the data set. An average level of four was set as the lower limit. Additionally, to minimise noise from participants who do not interact with others for a finite amount of time (e.g. because they have left campus or spend time with people not participating in the study), we filter the participants by their average node degree in the recent past:

$$\bar{k}_i(t) = \frac{\sum\limits_{t'=0}^{t} k_i(t') \cdot e^{-(t-t')^2/(2t_k^2)}}{\sum\limits_{t'=0}^{t} e^{-(t-t')^2/(2t_k^2)}} , \tag{2}$$

where k_i is the nodal degree and we have chosen as weight a one-sided Gaussian kernel $e^{-(t-t')^2/(2t_k^2)}$ with a characteristic time of $t_k = 7\,$days. Hence, the average $\bar{k}_i(t)$ can be understood as the number of contact events in approximately the last week. We set the lower bound to $\bar{k}_i(t) = 1/7$, which optimally minimises noise.

In order to investigate possible spreading dynamics of the illustrative behaviour "regularly going to the fitness studio", we match stop-locations with the locations of fitness studios (Appendix A). Here, stop-locations are coordinates generated from the GPS data, where the participants spent at least 15 minutes [62]. The accuracy chosen for matching is $10\,$m, which corresponds to the precision of GPS [63]. Hence, we record for each node i at time t the behaviour

$$b_i(t) = \begin{cases} 1\ , & \text{if node } i \text{ visited a studio at day } t \\ 0\ , & \text{otherwise} \end{cases} . \tag{3}$$

To distinguish between students who go to the studio occasionally and students who go regularly, we introduce the smoothed behaviour

$$\bar{b}_i(t) = \sum_{t'=0}^{t} b_i(t') \cdot e^{-(t-t')^2/(2t_b^2)} , \tag{4}$$

with the characteristic time $t_b = 7\,$days. The one-sided Gaussian kernel $e^{-(t-t')^2/(2t_b^2)}$ is chosen to favour current behaviour occuring close to time t, where the kernel reaches values close to one. Conversely, it suppresses past behaviour. Thus, $\bar{b}_i(t)$ can be interpreted as the typical behaviour in the last seven days.

Finally, for each point in time t we split the participants into two groups: (i) students going occasionally or not at all to the fitness studio $\bar{b}_i(t) < \gamma$ and (ii) students going regularly to the studio $\bar{b}_i(t) \geq \gamma$ and generate a time-dependent trait $o_i(t)$ for each node in the network,

$$o_i(t) = \begin{cases} 1 \text{ , } \bar{b}_i(t) \geq \gamma \\ 0 \text{ , otherwise} \end{cases} . \tag{5}$$

As threshold, $\gamma = 1$ is chosen, motivated by a clear edge in the cumulative distribution of $\bar{b}(t)$ plotted in Fig. 2. The edge is visible at $\bar{b}(t) \approx 1$ for all t, with values of $\bar{b}(t) > 1$ occurring less frequently than $\bar{b}(t) < 1$. This suggests that it is reasonable to separate participants between those who go to gyms regularly $\gamma \geq 1$, and those who go only occasionally $\gamma \leq 1$. The former will be referred to as "active" nodes, and the latter as "passive" nodes.

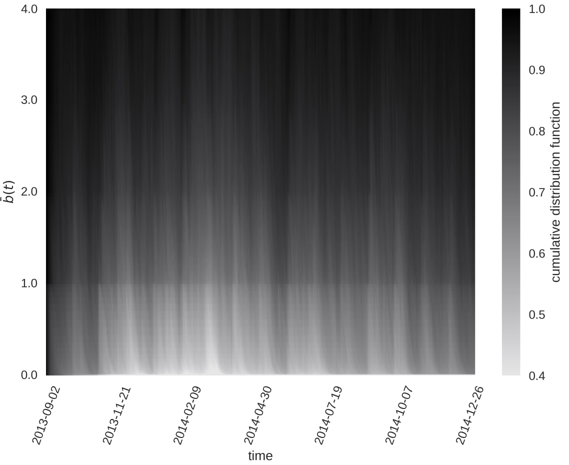


Fig. 2: Cumulative distribution of the smoothed behavioural function $\bar{b}(t)$ plotted as a heat map over the period of the entire "SensibleDTU 2013" data collection. Our study analyses the three month subperiod from February to April 2014. A clear edge is visible at $\bar{b}(t) \approx 1$ for all t, with values of $\bar{b}(t) > 1$ being much less frequent than $\bar{b}(t) < 1$. Therefore, $\gamma = 1$ is a reasonable choice to separate the participants into two groups. Members of the group with $\bar{b}(t) \geq 1$, who visit the fitness studio at frequent intervals, are referred to as active nodes, while individuals with $\bar{b}(t) < 1$ are referred to as passive nodes.

3 Methods

In this section, we describe the methodologies used to estimate empirical dose response functions from temporal network data (Sect. 3.1) and for generating surrogate data sets to test hypothesis on the processes and structures underlying specific features of the empirical dose response functions (Sect. 3.2).

3.1 Estimating dose-response functions from temporal network data

Dose response functions (DRFs) represent the functional dependence between the probability of changing a trait $p_{o \to o'}$ and the exposure K, which is defined as the joint influence of all contacts with a given trait, or more formally as the superposition of all received doses from neighbouring nodes. To measure the exposure to which a single node i is subjected, we put

$$K_i(o,t) = \sum_{t'=0}^{t} \mathcal{N}_i(o,t') \cdot e^{-(t-t')^2/(2t_K^2)} \text{ ,} \tag{6}$$

where $\mathcal{N}_i(o,t')$ is the number of neighbouring nodes with trait o at time t'. Hence, we assume that each node's influence is equal. The one-sided Gaussian kernel $e^{-(t-t')^2/(2t_K^2)}$ together with the characteristic exposure time of $t_K = 7$ days acts as a smoothing. Contacts in the near past $t - t' \lesssim t_K$ dominate the sum due to the weighting by the kernel. Conversely, contacts in the distant past $t - t' \gtrsim t_K$ are devalued. We thus interpret the kernel as representing the memory capacity of node i for the period of approximately $t_K = 7$ days.

From the time series of each node's traits $o_i(t)$, the received exposures $K_i(o,t)$ can be computed, allowing us to estimate the DRFs as relative frequencies as

$$p_{o\to o'}(K) \approx \frac{C(K)}{N(K)} \, . \tag{7}$$

Here $C(K)$ is the number of nodes that have changed their trait between $t-1$ and t and having experienced a certain level of exposure K. Furthermore, $N(K)$ is the total number of nodes that have experienced exposure level K. $C(K)$ and $N(K)$ are the result of an aggregation over all time steps and are thus time-independent.

$p(K)$ is an estimator of the actual probability of changing trait when experiencing an exposure level of K. If the reactions (changing trait or not) to subsequent exposures are assumed to be independent, this estimator is simply the empirical success rate of an N(K) times repeated Bernoulli experiment, and its standard error can thus be estimated by

$$\sigma_p = \frac{\sqrt{C(K)(N(K)-C(K))}}{N(K)} \, . \tag{8}$$

In the present study we adopted $\sigma_p^c = \sqrt{C(K)(N(K)+C(K))}/N(K)$ as a conservative upper bound to this error.

3.2 Generating surrogate data sets for hypothesis testing

To probe the empirical data from the Copenhagen Networks Study for contagion effects relating to the studied behaviour, we use the method of surrogate data sets. The surrogate data approach is a statistical method for identifying non-linearity, such as contagion effects, in time series. This is achieved by performing hypothesis tests on data sets that are generated from the empirical data by using Monte Carlo methods [64, 65]. Surrogate data sets have been used in the past to study a wide range of time series [66–68] and network data [69–71]. The method is described in the following paragraph, followed by the description of the surrogate data studies presented in this contribution.

First, a class of linear processes that may potentially be sufficient in explaining the empirical data, is specified as a composite null hypothesis $\mathcal{H}_0$. To test this hypothesis, a new, "surrogate" data set is derived from the empirical data in a way that is consistent with $\mathcal{H}_0$. Any potential non-linear features that the null-hypothesis excludes are destroyed in this process, while some linear features of the original data are retained. One algorithm which can be used to produce such surrogate data sets is the creation of random permutations of the original data. The product resembles the empirical data, but lacks any potential non-linearities, such as contagion processes. This method, known as Constrained Realisations [72], represents a parameter-free way of producing surrogate data sets without the use of a specific model. A discriminating statistic is then computed on the original data and surrogate data sets alike. If there is a significant difference between the value or distribution computed for the original data, and the ensemble of values or distributions computed for the surrogate data sets, the null hypothesis is rejected. Put simply, the empirical data are permuted in a way that is consistent with a composite null hypothesis, and if this substantially changes a statistical measure of interest, the null hypothesis can be rejected. Through the careful choice of iteratively more complex null hypotheses, preserving different sets of data properties, the nature of the true underlying non-linear process can be investigated.

Six surrogate data sets are produced for this analysis. The first four investigate the influence of different assumptions about the node dynamics on the dose response functions, by permuting the node traits $o_i(t)$ and keeping the network component $A_{ij}(t)$ unchanged. The last two surrogate data sets address the effect of the network component, by permuting the network edges $A_{ij}(t)$ and keeping the node dynamics $o_i(t)$ unchanged. In the following, the estimated DRF of the empirical data is referred to as the empirical DRF $p_{o\to o'}$, while the one estimated for surrogate data may be referred to as the surrogate DRF $\tilde{p}_{o\to o'}$. The following surrogate data test were conducted:

1. $\mathcal{H}_0^1$: *The empirical DRF can be reproduced with a class of models that is based only on the global mean activity level* $O = \overline{\langle o_i(t)\rangle_i}$. Here, the overline and brackets represent the time and ensemble average, respectively. This null hypothesis represents the most basic assumption, corresponding to an underlying process that is completely random. For this surrogate data set, all traits $o_i(t)$ are permuted randomly. Only the average activity level across the entire ensemble and observation period is conserved.
2. $\mathcal{H}_0^2$: *The empirical DRF can be reproduced with a class of models that is based only on each node's individual activity level* $O_i = \overline{o_i(t)}$. This null hypothesis leaves room for an activity factor unique to each individual node, while still assuming otherwise random node dynamics. For the corresponding surrogate data set, the activity levels are permuted in time, separately for each node.
3. $\mathcal{H}_0^3$: *The empirical DRF can be reproduced with a class of models that is based only on each node's individual activity level* O_i*, and its number of activity state switches.* This null hypothesis builds on the previous one by also conserving each node's persistence, defined as the inverse of a node's number of switches between behaviours. This is realised by separately permuting the length of intervals with a constant activity level, separately for periods of active and passive behaviour, for each node.
4. $\mathcal{H}_0^4$: *The empirical DRF can be reproduced with a class of models that is based only on the mean time-dependent activity level* $O(t) = \langle o_i(t)\rangle_i$ *of the ensemble.* This null hypothesis assumes a non-stationary temporal dynamics of the ensemble's behaviour, while excluding any non-random individual node characteristics. The surrogate data set is produced by permuting the activity states of all nodes, separately for each time step.

5. $\mathcal{H}_0^5$: *The empirical DRF can be reproduced with a class of models that is based only on individual activity dynamics and the average network edge density* $A = \overline{\langle A_{ij}(t) \rangle_{i,j}}$. In this case, the null hypothesis contains the assumption that the observed DRF is independent of the specific topology of the connection network, and arise solely based on the individual nodes' behaviour. The corresponding surrogate data set is produced by randomly permuting all edges across nodes and time.
6. $\mathcal{H}_0^6$: *The empirical DRF can be reproduced with a class of models that is based only on the individual node dynamics, and each node's time-dependent network degree* $k_i(t) = \sum_{j=0}^{N} A_{ij}(t)$. This null hypothesis builds on the previous one by randomising the neighbourhood of the nodes, but preserving each nodes connectivity in the network. This can serve as a check for homophilic effects in the network dynamics. To produce the surrogate data set, we use the random link switching algorithm [73, 74]. Pairs of connections (i, j) and (k, l) are drawn randomly, and are transformed into the connections (i, k) and (j, l). This procedure ensures that each node's degree remains unchanged.

We choose the dose response function, introduced in Sect. 3.1, as the discriminating statistic used to compare empirical and surrogate data sets. The comparisons of surrogate and empirical data sets are presented in Sect. 4.2.

4 Results

Here, we report on the results obtained by applying our proposed dose response function methodology. As a first step, we analyse synthetic data generated by the adaptive voter model as a proof of concept (Sect. 4.1). Building on these insights, we then investigate the empirical temporal network data obtained from the Copenhagen Network Study (Sect. 4.2).

4.1 Synthetic data

As a first application of our methodology, we analyse synthetic temporal network data generated by the adaptive voter model (Sect. 2.2). Fig. 3 shows the estimated DRFs for the AVM with $\varphi = 0$ (green dots), which includes only imitation dynamics, and with $\varphi = 0.6$ (blue crosses), involving both imitation and homophily dynamics. The plots contain the data from ten independent model runs each. The probabilities for the change of trait $p_{o \to o'}$ are generated for equally sized bins with a width of $K = 2$. Only bins with at least 30 data points were considered. Nevertheless, for high K, the DRF $p_{o \to o'}$ is subject to increasing uncertainties since exposures $K > 30$ are very rare in the network.

As suggested by the imitation rule in the model, we observe that $p_{o \to o'}$ depends monotonically, but non-linearly, on K. Moreover, the plot for $\varphi = 0.6$ clearly shows the impact on $p_{o \to o'}(K)$ of the additional homophily compared to the plot of $\varphi = 0$. For $K \gtrsim 15$ the DRF of this data is significantly larger then for those with $\varphi = 0$.

From this first proof of concept application, we can conclude that contagion dynamics such as the imitation rule in the model [2, 20] leads to positive correlation of $p_{o \to o'}$ and K. However, from the estimated DRF for $\varphi = 0.6$, we learned that homophily is reflected in the DRFs as well. To distinguish between the different dynamics, we use a surrogate analysis in the following investigation of the empirical temporal network data (Sect. 3.2).

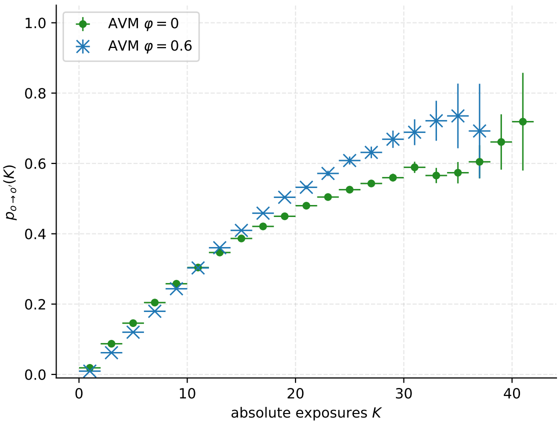


Fig. 3: Average estimated dose response functions (DRFs) for synthetic temporal network data generated by ten runs of the adaptive voter model for rewiring probability $\varphi = 0$ and $\varphi = 0.6$. The number of nodes $N = 850$ and the average degree $\langle k_i \rangle = 13.5$ were chosen analogously to the empirical temporal network from the Copenhagen Networks Study. The difference between the two DRFs shows that their form is not only influenced by contagion (imitation or social learning) effects, but also by homophily (network adaptation) dynamics.

4.2 Empirical data

In the following, we apply our methodology to empirical temporal network data from the Copenhagen Networks Study (Sect. 2.3) to investigate possible spreading dynamics of the illustrative behaviour "regularly going to the fitness studio". The DRF $p_{o\to o'}(K)$ is estimated for equal-sized bins with a width of $K = 5$. Only bins with at least 30 data points were considered. The resulting DRFs are shown in Fig. 4.

We observe that the probabilities for becoming active $p_{p\to a}$ (Fig. 4a) and for becoming passive $p_{a\to p}$ (Fig. 4b) do not behave in a symmetric way. Since the initiation and the maintenance of an activity represent two rather distinct phases [43], this is not necessarily surprising. For the latter, $p_{a\to p}$, a slight negative dependence on K may be indicated, however this is obscured by the large error bars. Stopping to regularly go to the fitness centre could possibly be largely independent of contagion events and dominated by external influences (e.g. an injury). Therefore, in the following we focus our analysis on the probability of becoming active $p_{p\to a}$.

The probability $p_{p\to a}$ is subject to large errors for $K > 100$. The low occurrence of large K seems to be the main reason. However, we find a notable positive correlation of $p_{p\to a}$ with K for $K < 100$, which could indicate contagion or homophilic dynamics. To pursue this indicator further, we examine the DRF using the surrogate data set method (Sect. 3.2). First, we investigate the possible influence of contagion dynamics (Sect. 4.2.1), then for group dynamics or external influences (Sect. 4.2.2) and finally for homophily dynamics (Sect. 4.2.3).

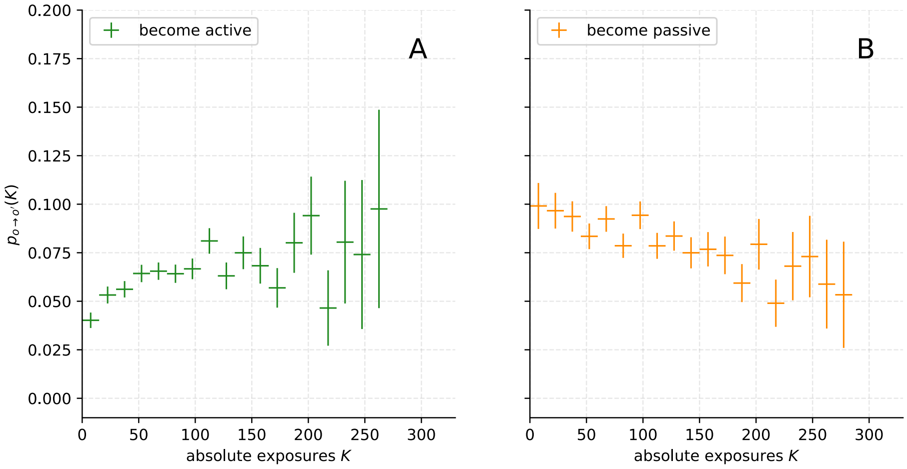


Fig. 4: Empirical dose response functions computed from the Copenhagen Networks Study temporal network data, representing the the probability to become active (A) or passive (B), as a function of the absolute exposure to these respective activity levels. For the probability to become active $p_{p\to a}$, a clear upward trend is noticeable, which might be caused by contagion, although the sparse data at $K > 170$ make it difficult to discern this trend there. For the probability to become passive $p_{a\to p}$, no clear dependence on K can be identified due to large uncertainties. Note that the two estimated DRFs are very different from those derived from the adaptive voter model shown in Fig. 3

.

4.2.1 Investigation for Contagion Dynamics

For investigating the possible influence of contagion dynamics on the DRF we employ the surrogate data tests $\mathcal{H}_0^1$, $\mathcal{H}_0^2$, and $\mathcal{H}_0^3$ introduced in Sect. 3.2, i.e., consider surrogate models in which explicitly no contagion takes place and we explore if they nevertheless reproduce the empirically observed DRF. To do so, we permute the traits of the nodes $o_i(t)$ and leave the network component $A_{ij}(t)$ unchanged. These permutations destroy possible temporal correlations of exposure K with changes in traits and, thus, any trace of contagion dynamics. In three steps, we analyse the impact of different assumptions about the node dynamics on the dose-response functions and show step by step which assumptions are necessary to explain the observed DRF.

First Data Test. Hypothesis $\mathcal{H}_0^1$: *The empirical DRF can be reproduced with a class of models that is based only on the global mean activity level $O = \overline{\langle o_i(t)\rangle}_i$.*

We test the most basic assumption of whether the empirical DRF can be explained by uncorrelated traits. To do so, all traits were uniformly permuted at random and only the global mean activity level $O = \overline{\langle o_i(t)\rangle_i}$, was conserved.

Here, the overline and the brackets represent the time and ensemble mean, respectively. All possible contagion dynamics are destroyed in the model due to the random permutations.
Expectation. We expect to observe no correlation between the DRF $\tilde{p}_{p\to a}$ of the surrogate and K due to the permutations. Moreover, $\tilde{p}_{p\to a}(K)$ should be equal to the fraction of active states in the whole observed period.
Result. In Fig. 5a, the DRF $\tilde{p}_{p\to a}$ of the surrogate is contrasted with the empirical DRF $p_{p\to a}$. We find our expectations confirmed, $\tilde{p}_{p\to a}$ is quantitatively and qualitatively different from $p_{p\to a}$. Moreover, $\tilde{p}_{p\to a}$ is approximately equal to the share of active states. Therefore, the model is not sufficient to explain the empirical dynamics and we reject the first null hypothesis.

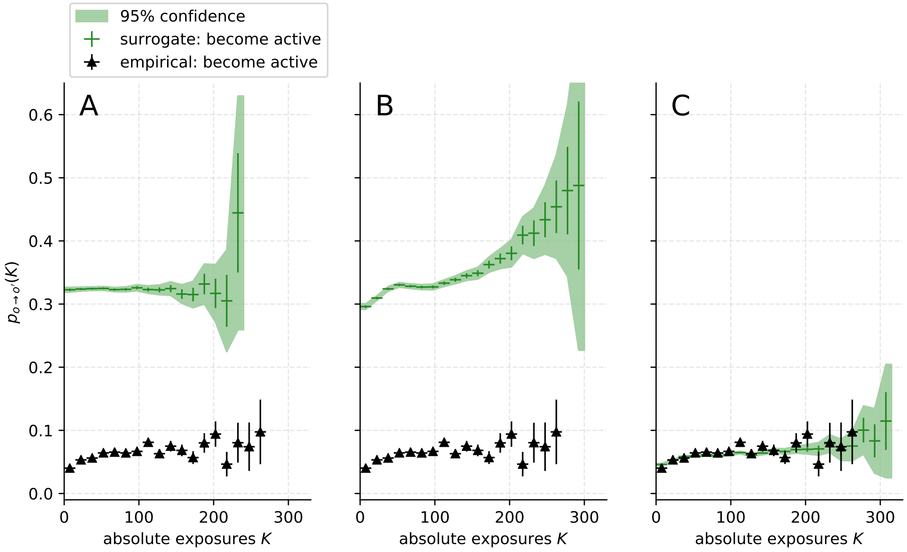


Fig. 5: Comparison of DRFs computed on empirical data (black triangles) and surrogates of the node traits (green crosses), corresponding to the null hypotheses $\mathcal{H}_0^1$ through $\mathcal{H}_0^3$. It can be observed that neither A) the preservation of the average trait O ($\mathcal{H}_0^1$), nor B) the additional preservation of each individual node's average trait O_i ($\mathcal{H}_0^2$) is sufficient to reproduce the data. C) However, when the individual node persistence, defined as the inverse of the number of trait switches, is also conserved ($\mathcal{H}_0^1$), the surrogate and empirical data show good agreement. Thus, we do not find sufficient evidence that contagion plays a significant role.

Second Data Test. Hypothesis $\mathcal{H}_0^2$: *The empirical DRF can be reproduced with a class of models that is based only on each node's individual activity level $O_i = \overline{o_i(t)}$.*

We test the effects of the individual activity level of each node $O_i = \overline{o_i(t)}$. Analogous to the previous model, the traits per node are randomly permuted in time, but this time not in the ensemble. Therefore, O_i is conserved. As in the previous model, any possible contagion dynamics are destroyed due to the permutations.
Expectation. Due to the permutation in the surrogate, the individual probability of the node to change its trait is equal to O_i. In particular, this probability is independent of the exposure K. Therefore, we do not expect any correlation between $\tilde{p}_{p\to a}$ and K.
Result. Contrary to our expectations, in Fig. 5b we find the probability $\tilde{p}_{p\to a}$ and K positively correlated, qualitatively similar to the correlation of $p_{p\to a}$ and K. However, for $K > 100$, the probability $\tilde{p}_{p\to a}(K)$ continues to increase, while $p_{p\to a}(K)$ appears to saturate. Furthermore, $\tilde{p}_{p\to a}$ and $p_{p\to a}$ differ quantitatively by a factor of about six. Thus, the conservation of O_i is not sufficient to explain the empirical DRF, and we also reject the second null hypothesis.

In the second considered model, we found that the DRFs of the surrogate and the empirical data behave in a qualitatively similar way. This could be the result of pre-existing clustering in the data set: contacts j of nodes i would have similar activity values $O_j \approx O_i$ over the entire observation period. A node i with e.g. low O_i thus has contacts j with low O_j and therefore receives low exposure K. A positive correlation would be the result. Even without fully understanding the cause of the correlation found, it can be concluded that the individual activity level O_i is an essential feature in the empirical network. In addition to the correlation, we found a shift of the DRF $\tilde{p}_{p\to a}(K)$ by a factor of six compared to $p_{p\to a}$. We suspect the reason for this shift to be the non-preserved persistence of the nodes (inverse number of individual activity state changes). Due to the random permutations, the nodes change their trait more frequently than in the empirical network. In the following surrogate, this hypothesis is analysed in more detail.

Reproduced from arXiv preprint: J.F. Donges, J. Lochner et al., Dose-response function approach for detecting spreading processes in temporal network data, arXiv:2103.09496. Published under arXiv.org - non-exclusive license to distribute.

Third Data Test. Hypothesis $\mathcal{H}_0^3$: *The empirical DRF can be reproduced with a class of models that is based only on each node's individual activity level O_i, and its individual persistence (inverse number of individual activity state switches).*

Additionally to O_i, the effect of individual persistence is tested. To achieve this, both the intervals with active trait $o_i(t) = 1$ and the intervals with passive trait $o_i(t) = 0$ were permuted at random. Hence, O_i and the persistence are conserved. Similar to the previous models, the random permutations remove any possible contagion dynamics.
Expectation. Due to the additional conservation of individual persistence, we expect $\tilde{p}_{p\to a}$ to be qualitatively similar to $\tilde{p}_{p\to a}$ from the second model, but shifted closer to the empirical DRF on the y axis.
Result. In Fig. 5c, we find, consistently with our expectations, that the DRF of the surrogate is shifted. Moreover, the probability $\tilde{p}_{p\to a}$ saturates for $K > 100$, analogous to the empirical DRF. Overall, no significant deviation between $\tilde{p}_{p\to a}$ and $p_{p\to a}$ can be found. Therefore, we cannot reject the third null hypothesis.

The third model showed that individual persistence is a main feature in the empirical network. Moreover, the model reproduces the empirical DRF in the model even without contagion. Thus, the third model shows that the data are not sufficient evidence that contagion plays a significant role in the empirical network, contrary to the hypothesis we formed when we first observed the correlation of $p_{p\to a}$ and K.

4.2.2 Investigation for Group Dynamics

In the previous section, we tested the effects of individual properties such as the individual activity level O_i or the individual persistence with our models. To investigate the importance of group dynamics, in this section we discard all individual properties and test the following null hypothesis:

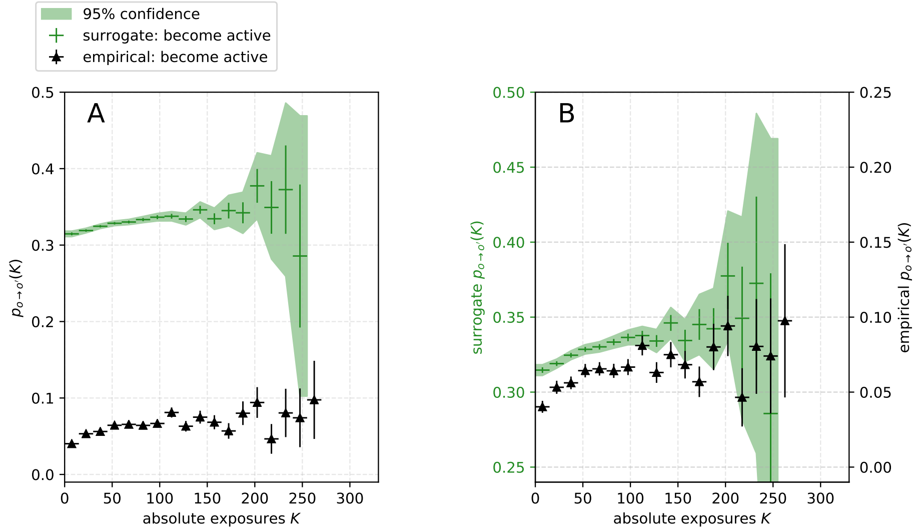


Fig. 6: Comparison of the DRF for empirical (black triangles) and surrogate (green crosses) data for null hypothesis ($\mathcal{H}_0^4$). To investigate external influences that affect all nodes simultaneously, the node traits were randomized in a way that conserves the time-varying mean activity level $O(t)$ of the group. The two figures contain the same data: A) compares the absolute values of the data points, while in B) the surrogate data y-axis (green, left side) is offset by 0.25 to facilitate comparison of the functional forms. While the absolute values differ strongly, similarities in the functional forms are apparent, pointing to the importance of external influences on the collective group dynamics.

Fourth Data Test. Hypothesis $\mathcal{H}_0^4$: *The empirical DRF can be reproduced with a class of models that is based only on the mean time-dependent activity level $O(t) = \langle o_i(t) \rangle_i$ of the ensemble.*

We test the relevance of the mean time-dependent activity level $O(t) = \langle o_i(t) \rangle_i$ for the empirical dynamics. To do this, the traits between nodes were permuted at random for each time point separately, and only $O(t)$ is preserved.
Expectation. Given the permutations, both the probability of becoming active $\tilde{p}_{p\to a}$ and the exposure K depend on $O(t)$. Thus, a correlation between $\tilde{p}_{p\to a}$ and K is to be expected. Furthermore, we expect $\tilde{p}_{p\to a}(K) \gg p_{p\to a}(K)$ resulting from the destruction of the persistence of the nodes.
Result. Fig. 6a compares the DRF $\tilde{p}_{p\to a}$ obtained from the surrogate data to the empirical DRF $p_{p\to a}$. Fig. 6b shows

Reproduced from arXiv preprint: J.F. Donges, J. Lochner et al., Dose-response function approach for detecting spreading processes in temporal network data, arXiv:2103.09496. Published under arXiv.org - non-exclusive license to distribute.

the same DRFs, but the DRF of the surrogate (green, left y-axis) is offset by 0.25 to better compare the shape of the functions. In line with our expectations, $\tilde{p}_{p\to a}$ is correlated with K. For $K < 100$, the probability $\tilde{p}_{p\to a}(K)$ increases linearly. The empirical $p_{p\to a}(K)$ also increases for $K < 100$, but slightly non-linearly. Quantitatively, we observe $\tilde{p}_{p\to a}(K) \gg p_{p\to a}(K)$. Thus, without individual traits, the model is not able to reproduce the empirical DRF. Therefore, we reject the fourth null hypothesis.

Although the surrogate model DRF is quantitatively significantly different from the empirical DRF, the model predicts a qualitatively similar functional form. Temporal group dynamics thus seems to be another important feature in the empirical temporal network data. Apparently, participants change their behaviour collectively, as is also evident from the fluctuations observed in the mean activity level (Fig. 2). Such non-stationarities could emerge from internal collective dynamics or be due to external influences such as, for example, exam periods, weekends or holidays. A more detailed analysis is needed to distinguish these possible effects.

4.2.3 Investigation for Homophily Dynamics

Continuing our investigation, we look for homophily dynamics in the network. Analogously to the analysis testing for contagion effects, we create surrogate models in which explicitly no homophily takes place. With these, we attempt to reproduce the empirical dynamics. To this end, we permute the network edges $A_{ij}(t)$ and keep the properties of the nodes $o_i(t)$ unchanged. This approach removes any homophily dynamics from the network, since the drawing and breaking of edges is randomised. The investigation is carried out in two steps, testing the following null hypotheses:

Fifth Data Test. Hypothesis $\mathcal{H}_0^5$: *The empirical DRF can be reproduced with a class of models that is based only on individual activity dynamics and the average network edge density $A = \overline{\langle A_{ij}(t)\rangle_{i,j}}$.*

We test the most basic assumption that the empirical dynamics can be explained by a random network. For this purpose, all edges were permuted uniformly at random. Only the average temporal network edge density $A = \overline{\langle A_{ij}(t)\rangle_{i,j}}$ was conserved. In this model, any homophily dynamics is removed, as the formation and breaking of edges is randomized.

Expectation. Since the traits have been kept unchanged, we expect the DRF of the model and the empirical DRF to be of the same order of magnitude. Due to the randomisation of the network, the neighbourhoods of the nodes are randomised as well. Thus, no correlation between the exposure K received from the neighbours and the probability $\tilde{p}_{p\to a}$ of changing the trait is to be expected.

Result. The DRF of the model and the empirical DRF are compared in the Fig. 7a. Contrary to our expectation, we can observe a correlation between $\tilde{p}_{p\to a}$ and K. Moreover, for the model, the case $\tilde{p}_{p\to a}(K)$ for $K > 100$ does not exist. Both DRFs have the same order of magnitude, which is in line with our expectations. However, only a few bins of the empirical DRF lie within the 95% confidence interval of the DRF from the surrogate. Consequently, we reject the fifth null hypothesis.

When analysing our model based on a random network, we observed a positive correlation between $\tilde{p}_{p\to a}$ and K. This correlation was significantly different from the correlation found for the empirical DRF. Therefore, the non-trivial network structure and dynamics appear to be essential for reproducing the empirical dynamics. One explanation for the correlation found could be the external influences already described in Sect. 4.2.2. Nodes may change their traits in synchrony, independently of the network and caused by an external influence. This would affect K as well and could explain the correlation found. A further analysis is necessary here. Another feature of the surrogate model's DRF is that no large exposure $K > 100$ occurred. This is likely caused by a much smaller variance of the degree distribution in the random network than in the empirical one. In the following surrogate, this hypothesis is analysed in more detail.

Sixth Data Test. Hypothesis $\mathcal{H}_0^6$: *The empirical DRF can be reproduced with a class of models that is based only on the individual node dynamics, and each node's time-dependent network degree $k_i(t) = \sum_{j=0}^{N} A_{ij}(t)$.*

Building on the previous model we test whether the time-dependent network degree of the nodes $k_i(t) = \sum_{j=0}^{N} A_{ij}(t)$ has a significant impact on the network dynamics. For this purpose, the edges of the network are permuted at random, but $k_i(t)$ is preserved. To generate the surrogate data set, we use the random link switching algorithm as described in Sect. 3.2. Analogous to the previous model, the homophily dynamics is removed by the permutations.

Expectation. For the correlation of $\tilde{p}_{p\to a}$ and K we expect it to be similar to the one of the previous model. However, for this model we conserved the node's degree. Thus, the progression of the DRF should also extend over $K > 100$.

Result. In Fig. 7b we compare the DRF of the model with the empirical one. In agreement with our expectation, we find $\tilde{p}_{p\to a}(K)$ for $K > 100$. However, the correlation of $\tilde{p}_{p\to a}$ and K is different from the previous model (Fig. 7a). No significant difference to the empirical DRF can be found anymore. Therefore, we cannot reject the sixth null hypothesis.

With this final surrogate model, we were able to reproduce the empirical DRF by conserving the node degree sequence in the temporal network data. Accordingly, node degree $k_i(t)$, the number of social contacts a student has at a given time time t within the student population covered by the study, seems to be an important feature in the empirical data set. Furthermore, the reproduction succeeded without including the dynamics of homophily. This shows that the empirical data provide not only no sufficient evidence for a significant influence of contagion (see the results for $\mathcal{H}_0^3$ reported above), but are also not sufficient evidence for a significant influence of homophily either.

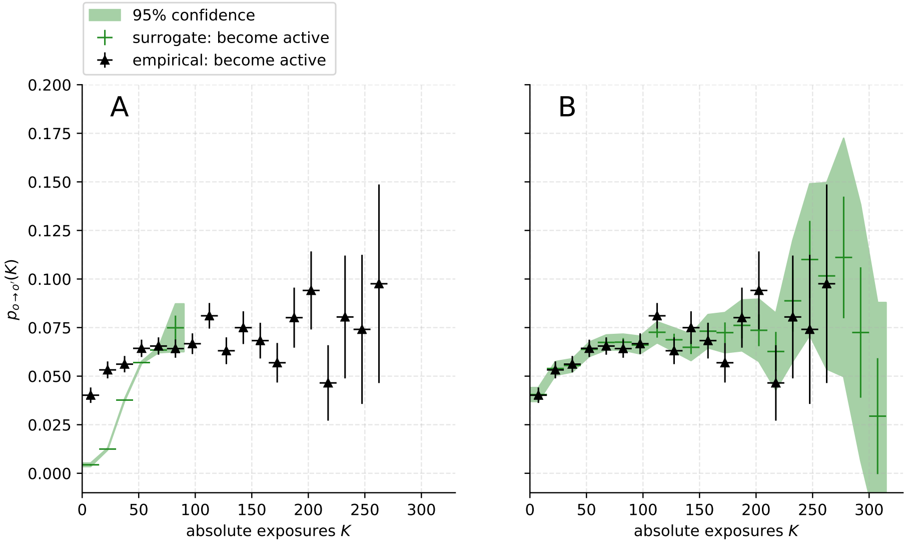


Fig. 7: Comparison of DRFs computed on empirical data (black triangles) and surrogates of the network topology (green crosses) for null hypotheses ($\mathcal{H}_0^5$) and ($\mathcal{H}_0^6$). In A) only the mean node degree k is conserved ($\mathcal{H}_0^5$), leading to a significant difference between empirical and surrogate data. In B) each node's time-varying degree $k_i(t)$ is conserved as well ($\mathcal{H}_0^6$), corresponding to a test for homophily in the network, with good agreement between the DRFs. It can be concluded that, while the non-trivial network structure appears to be of importance, no significant evidence for homophilic dynamics can be found.

5 Discussion and Conclusion

In this paper, we proposed a methodology for estimating dose response functions (DRFs) from temporal network data. We developed a hierarchy of surrogate data models to evaluate to what degree the observed DRFs can be explained by underlying processes such as social contagion, collective group dynamics and homophily. These surrogate models test the effects of distinct data features, such as overall and individual node activity levels, individual nodal trait persistence, overall network link density and individual node degrees. We applied this methodology to empirical temporal network data from the Copenhagen Networks Study, focusing on the illustrative health-related behaviour "regularly going to the fitness studio" in a physically-close-contact network of 850 university students, observed over the course of three months. The empirical data neither provide significant evidence for an influence of contagion, nor significant evidence for homophily. The individual activity level, individual behavioural persistence, effects of possibly externally forced collective group dynamics, and individual number of social contacts (the node degree sequence) are sufficient to explain the estimated empirical dose response function.

In the context of the application case considered in our study, these findings contradict the perspective that social interactions influence adopted behaviour, for example via subjective norms [75], as supported by psychological research [76]. In particular, the ability of social norms to influence individual decision-making has been identified previously as a potential tool for large-scale group behaviour transformations [11, 77]. However, in the present context of exercise behaviour a person may only be susceptible to social influence during particular stages of their decision process, while being almost "immune" at other times [43, 78]. At any time, too few people may be in this socially susceptible state to rise above the noise threshold in the data.

Overall, our results demonstrate that care needs to be taken in interpreting dose response functions obtained from empirical temporal network data; in particular when considering observational data that did not emerge from controlled experiments as in [36, 37]. Even pronounced positive correlations between exposure to a trait and the probability to adopt this trait can arise from structures in the temporal network data that do not need to be related to contagion and spreading processes, or homophily. Applying and further developing methodologies based on hierarchies of surrogate models, such as the one proposed in this article, provides a way forward to discern the specific imprints of complex spreading processes in temporal network data. Cases where the presence of such processes is not supported by the data can thus be excluded.

Our analysis has limitations in several dimensions that should be considered. Firstly, in terms of data limitations, the empirical temporal network data set extracted from the Copenhagen Networks Study depends on multiple assumptions on thresholds and other parameter values. The definition of social contacts as links in a physically-close-contact network could be too unspecific for discerning social contagion effects. Social contagion might be expected to require a more permanent and intense social relationship such as friendship to be effective. Furthermore, the definition of node traits as active or passive may suffer from noise and missing data issues, since most likely some fitness studios and other relevant exercise institutions (e.g. university gyms, swimming pools etc.) are missing from our list. Also, using GPS coordinates to determine whether a student is visiting a fitness studio introduces uncertainties: in a densely populated urban area like the city of Copenhagen, a café or a library might be located right next to, or even above or below a fitness studio, introducing additional noise into our data set.

Secondly, considering methodological limitations, DRFs are a highly aggregate statistical indicator describing a complex temporal network data set. They might not be specific enough to detect subtle spreading processes or to discriminate different types of complex contagions. Arguably this calls for higher order statistics with larger statistical power. Moreover, the proposed methodology based on a hierarchy of surrogate data sets is limited in that it allows only for indirect inference on the possible presence of spreading or contagion processes. In this respect it is desirable to augment the present analysis with more direct investigations including generative models of complex network spreading processes.

In summary, we suggest that our methodology is promising for applications to other systems and temporal network data sets. This can, among other applications, possibly aid our understanding of the social dynamics, spreading potentials and possible social tipping points in behaviours and social norms relevant for the adoption of healthy and sustainable diets [79] that can help to feed the world within planetary boundaries [80]. Efforts should be directed towards providing high-quality empirical temporal network data sets that can be leveraged for understanding complex spreading processes in these relevant domains. Promising directions of methodological developments include higher order statistics such as multi-node correlations for discerning the effects of longer contagion chains, spreading contagion waves, or the imprints of network motifs on complex spreading processes. Astute surrogate data models can provide detailed insights into such spreading processes. Connecting empirical network data to generative statistical and dynamical adaptive network models more directly, e.g. via maximum likelihood methods, appears similarly promising. Hence, one can open new perspectives to predict future spreading dynamics. Ultimately, this research thus aids in designing targeted interventions for fostering desirable or suppressing unwanted contagions in diverse complex systems including pandemics, brain, traffic and sustainability transformations.

Acknowledgements

The authors would like to thank Dr. Franziska Gutmann and Dr. Michaela Schinkoeth of the Sport and Exercise Psychology research group at University of Potsdam for a helpful discussion. JFD, JH and MW are thankful for financial support by the Leibniz Association (project DominoES). JFD acknowledges support from the European Research Council project Earth Resilience in the Anthropocene (743080 ERA). NK is grateful to the Geo.X Young Academy for financial support. SL acknowledges support by the Danish Research Council and the Villum Foundation.

References

1. Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.
2. Peter Sheridan Dodds and Duncan J. Watts. Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92(21):1–4, May 2004.
3. Sune Lehmann and Yong-Yeol Ahn. *Complex spreading phenomena in social systems*. Springer, 2018.
4. J D Murray. *Mathematical Biology : I . An Introduction*. Springer-Verlag, 2002.
5. D. J. Daley and J. Gani. *Epidemic Modelling*. Cambridge University Press, feb 1999.
6. Benjamin F Maier and Dirk Brockmann. Effective containment explains subexponential growth in recent confirmed COVID-19 cases in China. *Science*, 368(6492):742–746, 2020.
7. Sergey V Buldyrev, Roni Parshani, Gerald Paul, H Eugene Stanley, and Shlomo Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 2010.
8. James Samuel Coleman, Elihu Katz, and Herbert Menzel. *Medical innovation: A diffusion study*. Bobbs-Merrill Co, 1966.
9. ThomasW. Valente. Network models of the diffusion of innovations. *Computational and Mathematical Organization Theory*, 2(2):163–164, 1996.
10. Frank W Geels, Benjamin K Sovacool, Tim Schwanen, and Steve Sorrell. Sociotechnical transitions for deep decarbonization. *Science*, 357(6357):1242–1244, 2017.
11. Karine Nyborg, John M Anderies, Astrid Dannenberg, Therese Lindahl, Caroline Schill, Maja Schlüter, W Neil Adger, Kenneth J Arrow, Scott Barrett, Stephen Carpenter, et al. Social norms as solutions. *Science*, 354(6308):42–43, 2016.
12. J David Tàbara, Niki Frantzeskaki, Katharina Hölscher, Simona Pedde, Kasper Kok, Francesco Lamperti, Jens H Christensen, Jill Jäger, and Pam Berry. Positive tipping points in a rapidly warming world. *Current Opinion in Environmental Sustainability*, 31:120–129, 2018.

13. J Doyne Farmer, Cameron Hepburn, Matthew C Ives, T Hale, Thomas Wetzer, Penny Mealy, Ryan Rafaty, Sugandha Srivastav, and Rupert Way. Sensitive intervention points in the post-carbon transition. *Science*, 364(6436):132–134, 2019.
14. Ilona M Otto, Jonathan F Donges, Roger Cremades, Avit Bhowmik, Richard J Hewitt, Wolfgang Lucht, Johan Rockström, Franziska Allerberger, Mark McCaffrey, Sylvanus SP Doe, et al. Social tipping dynamics for stabilizing earth's climate by 2050. *Proceedings of the National Academy of Sciences*, 117(5):2354–2365, 2020.
15. Simon Sharpe and Timothy M Lenton. Upward-scaling tipping cascades to meet climate goals: plausible grounds for hope. *Climate Policy*, pages 1–13, 2021.
16. Susanne Lohmann. The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989–91. *World Politics*, 47(1):42–101, oct 1994.
17. Rodney Stark. Why religious movements succeed or fail: A revised general model. *Journal of Contemporary Religion*, 11(2):133–146, May 1996.
18. Robert L Montgomery. *The diffusion of religions: A sociological perspective*. University Press of America, 1996.
19. Ricarda Winkelmann, Jonathan F Donges, E Keith Smith, Manjana Milkoreit, Christina Eder, Jobst Heitzig, Alexia Katsanidou, Marc Wiedermann, Nico Wunderling, and Timothy M Lenton. Social tipping processes for sustainability: An analytical framework. *arXiv preprint arXiv:2010.04488*, 2020.
20. Peter S. Dodds and Duncan J. Watts. A generalized model of social and biological contagion. *Journal of Theoretical Biology*, 232(4):587–604, 2005.
21. Marc Wiedermann, E Keith Smith, Jobst Heitzig, and Jonathan F Donges. A network-based microfoundation of granovetter's threshold model for social tipping. *Scientific Reports*, 10(1):1–10, 2020.
22. Petter Holme and Mark EJ Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74(5):056108, 2006.
23. Thilo Gross, Carlos J Dommar D'Lima, and Bernd Blasius. Epidemic dynamics on an adaptive network. *Physical Review Letters*, 96(20):208701, 2006.
24. Thilo Gross and Hiroki Sayama. *Adaptive networks*. Springer, 2009.
25. Marc Wiedermann, Jonathan F Donges, Jobst Heitzig, Wolfgang Lucht, and Jürgen Kurths. Macroscopic description of complex adaptive networks coevolving with dynamic node states. *Physical Review E*, 91(5):052801, 2015.
26. Solomon Hsiang, Daniel Allen, Sébastien Annan-Phan, Kendon Bell, Ian Bolliger, Trinetta Chong, Hannah Druckenmiller, Luna Yue Huang, Andrew Hultgren, Emma Krasovich, et al. The effect of large-scale anti-contagion policies on the covid-19 pandemic. *Nature*, 584(7820):262–267, 2020.
27. Frank Schlosser, Benjamin F Maier, Olivia Jack, David Hinrichs, Adrian Zachariae, and Dirk Brockmann. Covid-19 lockdown induces disease-mitigating structural changes in mobility networks. *Proceedings of the National Academy of Sciences*, 117(52):32883–32890, 2020.
28. Peter J Menck, Jobst Heitzig, Jürgen Kurths, and Hans Joachim Schellnhuber. How dead ends undermine power grid stability. *Nature Communications*, 5(1):1–8, 2014.
29. Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, jul 2007.
30. Nicholas A. Christakis and James H. Fowler. The Collective Dynamics of Smoking in a Large Social Network. *New England Journal of Medicine*, 358(21):2249–2258, May 2008.
31. James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ (Online)*, 337(a2338), dec 2008.
32. John T. Cacioppo, James H. Fowler, and Nicholas A. Christakis. Alone in the Crowd: The Structure and Spread of Loneliness in a Large Social Network. *Journal of Personality and Social Psychology*, 97(6):977–991, 2009.
33. J. Niels Rosenquist, Joanne Murabito, James H. Fowler, and Nicholas A. Christakis. The spread of alcohol consumption behavior in a large social network. *Annals of Internal Medicine*, 152(7):426–433, apr 2010.
34. J N Rosenquist, J H Fowler, and N A Christakis. Social network determinants of depression. *Molecular Psychiatry*, 16(3):273–281, mar 2011.
35. Rose McDermott, James H. Fowler, and Nicholas A. Christakis. Breaking up is hard to do, unless everyone else is doing it too: Social network effects on divorce in a longitudinal sample. *Social Forces*, 92(2):491–519, dec 2013.
36. Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.
37. Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.
38. Elizabeth L Ogburn. Challenges to estimating contagion effects from observational data. In *Complex Spreading Phenomena in Social Systems*, pages 47–64. Springer, 2018.
39. Jakob Runge, Vladimir Petoukhov, Jonathan F Donges, Jaroslav Hlinka, Nikola Jajcay, Martin Vejmelka, David Hartman, Norbert Marwan, Milan Paluš, and Jürgen Kurths. Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature communications*, 6(1):1–10, 2015.
40. Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018.
41. Matthew MG Sosna, Colin R Twomey, Joseph Bak-Coleman, Winnie Poel, Bryan C Daniels, Pawel Romanczuk, and Iain D Couzin. Individual and collective encoding of risk in animal groups. *Proceedings of the National*

Academy of Sciences, 116(41):20556–20561, 2019.
42. Nathan O. Hodas and Kristina Lerman. The Simple Rules of Social Contagion. *Scientific Reports*, 4(1):4343, March 2014.
43. B. H. Marcus and L. R. Simkin. The transtheoretical model: applications to exercise behavior. *Medicine and Science in Sports and Exercise*, 26(11):1400–1404, November 1994.
44. Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012.
45. Marián Boguná, Romualdo Pastor-Satorras, Albert Díaz-Guilera, and Alex Arenas. Models of social networks based on social distance attachment. *Physical review E*, 70(5):056122, 2004.
46. Claudio Castellano, Daniele Vilone, and Alessandro Vespignani. Incomplete ordering of the voter model on small-world networks. *EPL*, 63(1):153, 2003.
47. Richard A. Holley and Thomas M. Liggett. Ergodic Theorems for Weakly Interacting Infinite Systems and the Voter Model. *The Annals of Probability*, 3(4):643 – 663, 1975.
48. Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. Measuring large-scale social networks with high resolution. *PloS one*, 9(4):e95978, 2014.
49. Piotr Sapiezynski, Arkadiusz Stopczynski, David Dreyer Lassen, and Sune Lehmann. Interaction data from the copenhagen networks study. *Scientific Data*, 6(1):1–10, 2019.
50. Enys Mones, Arkadiusz Stopczynski, Alex 'Sandy' Pentland, Nathaniel Hupert, and Sune Lehmann. Optimizing targeted vaccination across cyber–physical networks: an empirically based mathematical simulation study. *Journal of The Royal Society Interface*, 15(138):20170783, 2018.
51. Arkadiusz Stopczynski, Sune Lehmann, et al. How physical proximity shapes complex social networks. *Scientific reports*, 8(1):1–10, 2018.
52. Sadamori Kojaku, Laurent Hébert-Dufresne, Enys Mones, Sune Lehmann, and Yong-Yeol Ahn. The effectiveness of backward contact tracing in networks. *Nature Physics*, pages 1–7, 2021.
53. Laura Alessandretti, Piotr Sapiezynski, Vedran Sekara, Sune Lehmann, and Andrea Baronchelli. Evidence for a conserved quantity in human mobility. *Nature human behaviour*, 2(7):485–491, 2018.
54. Laura Alessandretti, Ulf Aslak, and Sune Lehmann. The scales of human mobility. *Nature*, 587(7834):402–407, 2020.
55. Vedran Sekara, Arkadiusz Stopczynski, and Sune Lehmann. Fundamental structures of dynamic social networks. *Proceedings of the national academy of sciences*, 113(36):9977–9982, 2016.
56. Anders Mollgaard, Ingo Zettler, Jesper Dammeyer, Mogens H Jensen, Sune Lehmann, and Joachim Mathiesen. Measure of node similarity in multilayer networks. *PloS one*, 11(6):e0157436, 2016.
57. Ioanna Psylla, Piotr Sapiezynski, Enys Mones, and Sune Lehmann. The role of gender in social network organization. *PloS one*, 12(12):e0189873, 2017.
58. Valentin Kassarnig, Andreas Bjerre-Nielsen, Enys Mones, Sune Lehmann, and David Dreyer Lassen. Class attendance, peer similarity, and academic performance in a large field study. *PloS one*, 12(11):e0187078, 2017.
59. Valentin Kassarnig, Enys Mones, Andreas Bjerre-Nielsen, Piotr Sapiezynski, David Dreyer Lassen, and Sune Lehmann. Academic performance and behavioral patterns. *EPJ Data Science*, 7(1):10, 2018.
60. OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2019.
61. Vedran Sekara and Sune Lehmann. The strength of friendship ties in proximity sensor data. *PloS one*, 9(7):e100915, 2014.
62. Andrea Cuttone, Jakob Eg Larsen, and Sune Lehmann. Inferring human mobility from sparse low accuracy mobile sensing data. In *UbiComp 2014 - Adjunct Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 995–1004, New York, NY, USA, sep 2014. Association for Computing Machinery, Inc.
63. United States Department Of Defense. Global positioning system standard positioning service performance standard. Technical Report 4th Edition, 2008.
64. James Theiler, Stephen Eubank, André Longtin, Bryan Galdrikian, and J. Doyne Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1-4):77–94, September 1992.
65. Thomas Schreiber and Andreas Schmitz. Surrogate time series. *Physica D: Nonlinear Phenomena*, 142(3-4):346–382, August 2000.
66. V. Venema, S. Bachner, H. W. Rust, and C. Simmer. Statistical characteristics of surrogate data based on geophysical measurements. *Nonlinear Processes in Geophysics*, 13(4):449–466, 2006.
67. José A. Scheinkman and Blake LeBaron. Nonlinear Dynamics and Stock Returns. *The Journal of Business*, 62(3):311–337, 1989.
68. Walter S. Pritchard, Dennis W. Duke, and Kelly K. Krieble. Dimensional analysis of resting human EEG II: Surrogate-data testing indicates nonlinearity but not low-dimensional chaos. *Psychophysiology*, 32(5):486–491, 1995.
69. Marc Wiedermann, Jonathan F. Donges, Jürgen Kurths, and Reik V. Donner. Spatial network surrogates for disentangling complex system structure from spatial embedding of nodes. *Physical Review E*, 93(4):042308, April 2016.
70. Sergei Maslov, Kim Sneppen, and Alexei Zaliznyak. Detection of topological patterns in complex networks: correlation profile of the internet. *Physica A: Statistical Mechanics and its Applications*, 333:529–540, February 2004.

71. Sergei Maslov and Kim Sneppen. Specificity and Stability in Topology of Protein Networks. *Science*, 296(5569):910–913, May 2002.
72. James Theiler and Dean Prichard. Constrained-realization Monte-Carlo method for hypothesis testing. *Physica D: Nonlinear Phenomena*, 94(4):221–235, July 1996.
73. Gorka Zamora-López, Vinko Zlatić, Changsong Zhou, Hrvoje Štefančić, and Jürgen Kurths. Reciprocity of networks with degree correlations and arbitrary degree sequences. *Physical Review E*, 77(1):016106, January 2008.
74. Yael Artzy-Randrup and Lewi Stone. Generating uniformly distributed random networks. *Physical Review E*, 72(5):056708, November 2005.
75. Icek Ajzen. The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2):179–211, December 1991.
76. Albert Bandura. A social cognitive theory of personality. In *Handbook of personality*, pages 154–196. Guilford Publications, New York, 2nd edition, 1999.
77. H. Peyton Young. The Evolution of Social Norms. *Annual Review of Economics*, 7(1):359–387, 2015.
78. James O. Prochaska and Bess H. Marcus. The transtheoretical model: Applications to exercise. In *Advances in exercise adherence*, pages 161–180. Human Kinetics Publishers, Champaign, IL, England, 1994.
79. Walter Willett, Johan Rockström, Brent Loken, Marco Springmann, Tim Lang, Sonja Vermeulen, Tara Garnett, David Tilman, Fabrice DeClerck, Amanda Wood, et al. Food in the anthropocene: the eat–lancet commission on healthy diets from sustainable food systems. *The Lancet*, 393(10170):447–492, 2019.
80. Dieter Gerten, Vera Heck, Jonas Jägermeyr, Benjamin Leon Bodirsky, Ingo Fetzer, Mika Jalava, Matti Kummu, Wolfgang Lucht, Johan Rockström, Sibyll Schaphoff, et al. Feeding ten billion people is possible within four terrestrial planetary boundaries. *Nature Sustainability*, 3(3):200–208, 2020.

A List of considered fitness centers in Copenhagen

Name	Longitude [° E]	Latitude [° N]
Fresh Fitness Hvidovre	12.4691961	55.6415696
Fitness.dk	12.5618214	55.6614733
FitnessDK	12.5114098	55.6647699
Fresh Fitness	12.5404751	55.6975516
Fresh	12.4199488	55.6493081
Fitness World	12.4418141	55.7231967
Fitness World Ballerup	12.3579672	55.7296181
Fitness World Brøndby	12.4383494	55.6673030
Fitness World Farum Park	12.3513120	55.8172970
Fitness World Frederiksberg Bernhard Bangs Alle	12.5104671	55.6844058
Fitness World Frederiksberg Forum	12.5524718	55.6830906
Fitness World Frederiksberg Peter Bangs Vej	12.5131680	55.6795400
Fitness World Gentofte	12.5378949	55.7386120
Fitness World Glostrup	12.4008395	55.6640800
Fitness World Greve Hundige Storcenter	12.3274148	55.5987709
Fitness World Greve	12.2984612	55.5905648
Fitness World Herlev	12.4160534	55.7253403
Fitness World Husum	12.4810239	55.7095419
Fitness World København Baron Boltens Gård	12.5848511	55.6820125
Fitness World København Ellebjergvej	12.5108247	55.6507568
Fitness World København Emdrup Station	12.5409464	55.7218740
Fitness World København Englandsvej	12.6043943	55.6569690
Fitness World København Gasværksvej	12.5570237	55.6708078
Fitness World København Jagtvej	12.5509410	55.6964980
Fitness World København Lyngbyvej	12.5604444	55.7116463
Fitness World København Lyongade	12.6099453	55.6613686
Fitness World København Nordre Fasanvej	12.5364747	55.6985181
Fitness World København Strandvejen	12.5777058	55.7219712
Fitness World København Vester Farimagsgade	12.5623173	55.6782088
Fitness World København Århusgade	12.5872772	55.7067752
Fitness World Lyngby	12.5039072	55.7688801
Fitness World Måløv	12.3187172	55.7485909
Fitness World Søborg	12.4932893	55.7395909
Fitness World Taastrup	12.3017208	55.6529634
Fitness World Valby Mosedalvej	12.5134815	55.6674858
Fitness World Værløse	12.3615021	55.7821745
fitnessdk	12.4392816	55.7249089

Table 1: List of the fitness centers in Copenhagen considered in this study, with their respective coordinates, as extracted from Open Street Maps [60].

4.2 *Special cases of socio-economic dynamics and social tipping*

THIS SECOND SECTION is dedicated to the presentation of selected scenarios in which we identified socio-economic dynamics and social tipping.

Social tipping dynamics has been suggested as a key aspect of addressing the climate crisis [Otto et al., 2020b, Lenton, 2020, Farmer et al., 2019]. Dedicated minorities could encourage larger populations to get involved and fight global warming. In "A network-based microfoundation of Granovetter's threshold model for social tipping" [Wiedermann et al., 2020], we extended Granovetter's widely studied theoretical threshold model of collective behaviour to model social tipping phenomena on networks.

We continue in "Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure" [Schleussner, C. F. and Donges, J. F. et al., 2016] with an examination of large-scale transitions in societies. In this publication, we modeled the feedback effects of individual behavioural change and homophilic social network restructuring and illustrated their interplay by simulating how smoking behaviour and network structure are reconfigured by changing social norms.

In "Emergent inequality and endogenous dynamics in a simple behavioral macroeconomic model" [Asano et al., 2019, not included in this reader], we presented a simple macroeconomic model in which households are embedded in a social network. Unlike standard macroeconomic models, we did not assume that households base their decisions on utility maximization but are influenced by the behaviour of their neighbours via social learning. We could show that inequality and realistic business cycles both occur spontaneously as a consequence of imperfect household decision-making.

Divestment is seen as a key tool to achieve the goals of the Paris climate agreement [Rockström et al., 2017]. In "Divestment may burst the carbon bubble if investors' beliefs tip to anticipating strong future climate policy" [Ewers, B. and Donges, J. F. et al., 2019, not included in this reader], we presented an investigation of the dynamics of fossil fuel divestment using an adaptive social network coupled to a model of stock trading on a financial market. Our analysis highlights the potential for social tipping away from a fossil fuel-based economy.

We conclude this section with an analysis using a game-theoretic model of far-sighted coalition formation. In "Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility" [Heitzig and Kornek, 2018], we examined the dynamics of carbon market linkage among countries.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

natureresearch

Check for updates

OPEN

A network-based microfoundation of Granovetter's threshold model for social tipping

Marc Wiedermann[1]✉, E. Keith Smith[2,5], Jobst Heitzig[1] & Jonathan F. Donges[3,4]

Social tipping, where minorities trigger larger populations to engage in collective action, has been suggested as one key aspect in addressing contemporary global challenges. Here, we refine Granovetter's widely acknowledged theoretical threshold model of collective behavior as a numerical modelling tool for understanding social tipping processes and resolve issues that so far have hindered such applications. Based on real-world observations and social movement theory, we group the population into certain or potential actors, such that – in contrast to its original formulation – the model predicts non-trivial final shares of acting individuals. Then, we use a network cascade model to explain and analytically derive that previously hypothesized broad threshold distributions emerge if individuals become active via social interaction. Thus, through intuitive parameters and low dimensionality our refined model is adaptable to explain the likelihood of engaging in collective behavior where social-tipping-like processes emerge as saddle-node bifurcations and hysteresis.

Studies of collective behavior or action, such as protest demonstrations, responses to disasters or even revolution[1], fosters an understanding of the formation and logic of the *crowd*[2–5]. Broadly, the study of collective behavior can be separated into either that of *social movements* or that of *temporary gatherings*. Social movements are usually more structured around specific, identified goals, have deeper social connections between actors, are organized (generally to defend or fight against existing authorities) and persist over time (such as the civil rights movements)[6]. In contrast, gatherings (such as riots, sudden protests, concerts, sporting events) are more spontaneous, less organized, do not carry as deep of social connections between actors, and can be quite ephemeral[7,8].

Further, individual engagement in collective behaviors (such as changing consumption behavior or adoption of new technologies) can be connected to broader social processes, such as norms and expectations for behavior[9]. Specifically, individuals strategically control their actions in accordance with their norms in order to achieve their goals and objectives[4,5,10]. As such, norms and preferences structure an actor's likelihood to engage in collective behaviors, as well as its form of participation within these groups. Complex forms of collective behaviors (be it either a movement or a crowd) are thus created through dynamic interactions of actors that share common goals and objectives for a given social situation. For example, global climate change has been frequently noted as one prominent contemporary social problem that could trigger and might also be addressed through collective behaviour (such as the emergent 'Fridays for Future'[11] movement)[12–14].

Empirical evidence for such complex contagion of interlinked individuals leading to collective action has been found for both online[15–17] and offline[18] social networks. Additionally, complex contagion has been experimentally shown to foster social tipping[19], a process that has gained increased attention in the recently[20] due to its potential for rapid societal changes with profound impacts on the entire socio-ecological Earth System[13,21]. Complementing empirical studies, recent conceptual models of complex contagion incorporate the spreading of an action, behaviour or trait through a complex network[22–26]. They often aggregate an individual's surrounding over time[27,28] or abstract space[29] to accumulate exposure to a considered trait such that at a certain point the individual adopts that trait as well. Such models have been applied successfully to study processes involved in the

[1]FutureLab on Game Theory & Networks of Interacting Agents, Complexity Science, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, P.O. Box 60 12 03, 14412, Potsdam, Germany. [2]GESIS — Leibniz Institute for the Social Sciences, Member of the Leibniz Association, Unter Sachsenhausen 6-8, 50667, Cologne, Germany. [3]FutureLab Earth Resilience in the Anthropocene, Earth System Analysis, Potsdam Institute for Climate Impact Research, Member of the Leibniz Association, P.O. Box 60 12 03, 14412, Potsdam, Germany. [4]Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19, Stockholm, Sweden. [5]Institute of Science, Technology and Policy, ETH Zurich, Zurich, Switzerland. ✉e-mail: marcwie@pik-potsdam.de

spreading of opinions[30,31], large-scale epidemics[24], the adoption of life-style choices[32] or the collective behaviour of animal groups[33,34]. However, most such models of collective behavior are often tailored to a specific problem (both in the incorporated processes as well as the underlying parameter set) and are thus often not transferable to different and novel applications.

The *Granovetter threshold model* is a comparatively early contribution to this field, providing a core basis for subsequent and more contemporary modeling attempts[35]. This model aims to explain the emergence of collective behaviors while noting that individual norms and preferences are a crucial factor determining their development and final outcome. In particular, when presented with a simple binary choice – to participate within a collective behavior or not – each individual has a certain activation threshold for participation. This measures the proportion of the group that an individual would like to observe participating within the collective behavior before they are willing to join themselves. The thresholds emerge from the norms, preferences, goals and beliefs of each individual, e.g., representing a kind of trade-off between the costs and benefits of joining in the behavior. As such, the application of the threshold model, or variations thereof, is not limited to simple crowd-like behaviors, such as protests and riots, but is comparatively broad, encompassing collective behaviors e.g., voting[36], diffusion of innovations[37], or migration[38], as well as classical social movements such as the Monday Demonstrations in East Germany[39]. However, while by design the model is very flexible, it has mainly been used for illustrative and theoretical purposes (including most applications outlined above), but hardly applied as a numerical modeling tool.

This paper identifies two major sets of issues that prevent broader application of the Granovetter model and proposes extensions to resolve them. First, under often assumed threshold distributions (such as cut-off Gaussians[35]) the model usually unrealistically predicts either no-one or the entire population to eventually act. We resolve this issue by drawing from real-world observations, social movement and resource mobilization theories[40,41], as well as recent theoretical and numerical results regarding network spreading processes[42,43] to extend the original model by classifying individuals as either certainly active, certainly inactive, or contingently active. This causes the model to display nontrivial equilibria in which a certain part of the contingent individuals becomes active. Second, the emergence and shape of the threshold distribution itself is often underexplained. Therefore, we utilize an established conceptual network cascade model[29] and show that a broad (non-Gaussian) threshold distribution emerges from microscopic networked interactions in which potentially active individuals join an action if a sufficient number of their neighbors are also engaged. We thus specifically acknowledge empirically observed tendencies of individuals to make decisions with respect to their immediate social surrounding rather than considering the entire global population, i.e., the mean field[19,44,45]. By addressing both of the above issues, we effectively separate (unique) individual preferences which determine general tendencies towards or against an action from the embedding of each individual into a larger social structure and corresponding exposure to external influences. Both characteristics then co-determine whether the individual ultimately joins into an action or not.

The remainder of this work is organized as follows. We first introduce the formal specifics of the Granovetter threshold model and discusses in detail its aforementioned conceptual limitations. We then implement the proposed solutions and present a refined threshold model that only depends on parameters that are readily observable in real-world systems. Additionally, we provide an analytical solution of the refined model and analyse its potential for modeling social tipping. Ultimately, we culminate with a discussion of the results and an outlook to future work.

Granovetter's threshold model

The threshold model assigns each individual in a population of size N a threshold that defines the number of others that must participate in an action before the considered individual does so, too[35]. In its discrete-time formulation the number of acting individuals at time $t+1$, $R(t+1)$, is hence directly derived from the cumulative distribution function of thresholds in the population, F, such that

$$R(t+1) = NF(R(t)). \tag{1}$$

Note that the original exemplary application of the model was that of individuals' participation in riots. Hence the choice of the symbol R for the number of acting individuals. An equilibrium number of acting individuals R^* is obtained by solving $R(t+1) = R(t) = NF(R(t))$ for $R(t)$ which is equivalent to finding an intersection of the graph of F with the diagonal through $(0, 0)$ and (N, N), Fig. 1a. All equilibrium points R^* at which F intersects the diagonal line from above are stable, while all others are unstable[35].

While the threshold model has been widely used within a broad literature[41,46,47] it has up to now been mainly used for illustrative purposes as a number of issues hinder its application as numerical modeling tool:

Plausible distributions typically predict no one or the entire population to act.

As thresholds are hard to estimate, one typically assumes Gaussian threshold distributions[35] cut off at the extreme values 0 and N. However, assuming a mean threshold μ of reasonable size and a moderate standard deviation σ implies that there are only few individuals with low or high thresholds and many with medium thresholds close to μ. Hence, under the typical assumption of a low number of *instigators*[35] the model usually predicts zero eventually acting individuals, Fig. 1a. Only if a sufficiently large σ is chosen more individuals than the instigators become active. However, the choice of a large σ causes the distribution to become rather flat instead of bell-shaped. For example, for a population size of $N = 100$ and an average threshold of $\mu = 25$, a standard deviation of $\sigma = 12.2$ is required so that a single instigator can cause the rest of the population to become active[35].

In addition, if no individual has a threshold larger than 100%, the threshold model generally has a second typically stable fixed point at $R^* = N$ implying that the entire population has the potential to become active if only enough others do so, too, Fig. 1a. In reality, an individual may never engage in an action regardless of how

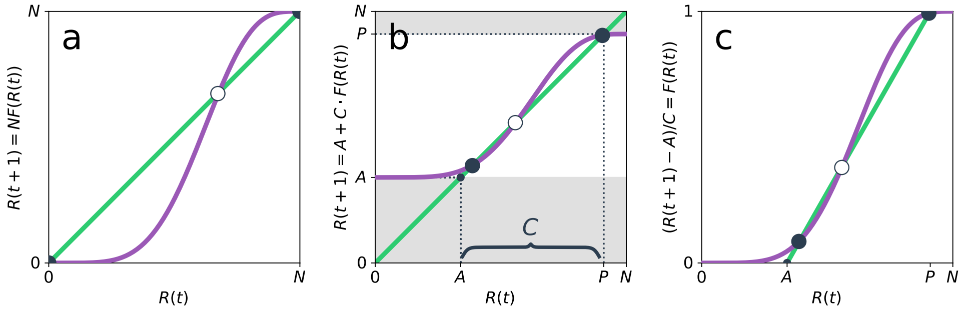


Figure 1. Extension of Granovetter's (graphic) model with P potentially and A certainly acting individuals. (**a**) The original model that computes the number of acting individuals $R(t+1)$ from the cumulative distribution function of thresholds F. The purple line indicates a typical normal-like choice for this distribution. The 45°-line (green) intersects F at the stable (black) and unstable (white) equilibrium points R^*. As for many realistic choices of F, only $R^*=0$ and $R^*=N$ are stable. (**b**) Introducing A certainly and P potentially acting individuals, such that the $C=P-A$ contingent individuals have the same threshold distribution F as the entire population N. Here, the equilibria move to the interval $R^* \in [A, P]$ and are not necessarily located at exactly $R^*=A$ and $R^*=P$. Hence, the A certainly acting individuals trigger some contingent individuals to act, too. (**c**) Rescaling $R(t+1)$ to the unit interval shows that equilibria can be computed by shifting the diagonal line from crossing $(0, 0)$ and (N, N) (as in (**a**)) to crossing $(A, 0)$ and $(P, 1)$ and using the same threshold distribution F as in (**a**).

many others have already joined as personal preferences, norms or attitudes can restrict behaviours[9]. In its basic setup, the Granovetter model can only account for this by either assigning the concerned individuals a threshold of 100% or by selecting the population such that only those individuals that are generally in favour of a certain action are considered[35]. The first approach, however, implies that everyone would generally be willing to act if only enough other individuals become active before. The second approach requires updating the population and, hence, its size, whenever the norms and attitudes of an individual change. What both approaches have in common is that they imply a constant change of the threshold distribution whenever individuals alter their preferences or attitudes.

We therefore propose a framework that refines the threshold model and accounts for the above issues by grouping individuals according to basic preferences that determine whether they certainly, contingently or never act. This circumvents the existence of trivial solutions and we show below that this approach does not require a constant updating of the threshold distribution as a response to changing group memberships.

The threshold distribution can not be observed, but emerges from microscopic factors.

Broadly, two complementary aspects shape whether an individual engages in an action or not. On the one hand there are *individual* factors (such as background characteristics, social class, education or occupation[48,49]), that determine the acceptance of or inclination towards an action. On the other hand there are *group* factors, i.e., characteristics resulting from one's embedding in a larger social network (such as social position, influence, or peer pressure[50]). Both traits and processes ultimately co-determine the macroscopic threshold that is exposed to the observer and we call these thresholds of the original Granovetter model *emergent thresholds* from here on. However, quantifying the emergent thresholds on the individual basis is difficult, if not impossible, to achieve without any prior knowledge or assumptions on the aforementioned microscopic characteristics and interactions. In addition, even properly justifying a certain shape of the emergent threshold distribution is a difficult task as it remains unclear to which extent different shapes follow from a certain composition of individual traits.

Notably, in analogy to the concept of emergent thresholds there should still exist on the micro-level a share (or number) of others that join into an action before an individual does so, too. One commonly accepted definition of such a quantity is that of a *threshold fraction*[29] that is not assessed with respect to the entire population, but with regard to the relevant social ties of a considered individual[35,51]. The specific importance of one's egocentric social network for decision making has recently been shown in empirical studies where individuals generally did not aim for consensus or convergence in the global population, but rather on the microscopic or group-level[19,44]. Additionally, it was observed that individuals tend to coordinate with (at least subsets of) an entire group rather a single partner[45]. This renders the use of a per-individual threshold fraction particularly useful as it determines the share of others within a group that must make a certain decision before the considered individual does so, too. In our specific case this threshold fraction is considered a fundamental trait of each individual, regardless of whether their preferences and norms favour or hinder a certain action. As such it disentangles social processes from non-social factors, such as individual preferences and norms. In contrast to the emergent thresholds, these threshold fractions may not necessarily be widespread. Rather, they might be assumed to have a narrow distribution or correspond to fixed, intuitive points, e.g. 50% (majority rule)[52]. Note that in contrast to the emergent thresholds, that measure *absolute* numbers in a global population, the threshold fraction measures the *relative* number of others in one's egocentric social network that must make a decision before a considered individual does so, too. It thereby specifically accounts for heterogeneities in the number of each individual's neighbors, i.e., the so-called social network's degree distribution[53].

Reproduced from: M. Wiedermann et al., A network-based microfoundation of Granovetter's threshold model for social tipping, Sci. Rep., vol. 10, no. 1, pp. 1-10, 2020, doi: 10.1038/s41598-020-67102-6. Published under Creative Commons Attribution License 4.0 (CC BY).

Below we present a microscopic threshold model based on a previous study of cascading dynamics[29] where individual preferences are assigned to each member of the population that then join into an action based on their threshold fractions applied to the neighborhood in their social network. We then show that such microscopic processes in fact yield an often postulated broad (but not normal-shaped) emergent threshold distribution.

Results

Refinement of the Model.

We start by addressing the first two issues identified above, namely that for usually chosen distributions the original model predicts either no-one or the entire population to become active. As discussed above, one way to circumvent these issues is to assign certain individuals either a threshold of 0% or ≥100% such that some individuals *certainly* become active and others never become active[35]. This approach requires a constant updating of the threshold distribution and may be impracticable for many cases. Recent studies investigated the effects of either such certainly active *initiators*[42] or never active *immune* individuals[43] on the adoption of certain traits or behaviours via spreading dynamics on social networks. In alignment with social movement theory[40,41] we combine these two notions and suggest to divide the population of size N into three groups, namely: $A \leq N$ *certainly acting* individuals[42], $C \leq N - A$ *contingent* individuals and the remaining $N - C - A$ *certainly inactive* individuals[43]. The certainly acting and contingent individuals form the group of $P = A + C$ *potentially acting* individuals. In a social movement and resource mobilization context, our three groups can for example be seen as representing adherents, potential supporters and those in opposition[40,41].

If we have no reason to assume that the threshold distribution is different in the three groups, the original recursive formula Eq. (1) is then replaced by

$$R(t+1) = A + C \cdot F(R(t)). \tag{2}$$

The equilibria of the thus refined model are again obtained by computing the intersection of the r.h.s. of Eq. (2) with the diagonal through $(0, 0)$ and (N, N), Fig. 1b. It is apparent that if $A > 0$ and $P < N$ (note again that $P = A + C$), we get nontrivial equilibrium numbers of acting individuals $R^* \in [A, P]$. Conveniently, as A or P (and C) change, the new equilibria can be found without re-estimating the threshold distribution.

In order to also avoid having to redraw F in Fig. 1b whenever there is a variation in A or C, it is beneficial to rescale the ordinate to the unit interval, Fig. 1c. This allows us to find the equilibria for all possible combinations of A and P in the same diagram, by drawing F only once and just adjusting the diagonal to meet the points $(A, 0)$ and $(P, 1)$.

Our adjusted approach makes the application of the threshold model as an actual modeling framework more practical as it (i) produces nontrivial fixed points R^*, (ii) requires the threshold distribution to be only estimated once for the entire population or a representative sample thereof, and (iii) relies on only two intuitive parameters, the size of the certainly (A) and potentially acting population (P). Recall that A directly relates to an immediate action or behaviour, while P denotes the general acceptance of or attitude towards that action.

Estimation of the emergent threshold distribution.

Having refined the threshold model to properly allow for the computation of non-trivial fixed points, we shift our focus to the second issue that relates to the threshold distribution itself. It has been established above that the emergent thresholds follow from microscopic characteristics of each individual as well as its embedding in a social context. Specifically for the latter it will turn out that the share of others, i.e., the threshold fraction, that must join into an action before a contingent individual does so, too need not be widely distributed or even heterogeneous at all across the population in order to produce a widespread distribution for the emergent threshold.

We now study how such characteristics and interactions on the micro-level determine one's emergent threshold by using a simulation model of social contagion that has been studied in the past to model binary decisions with externalities and resulting cascading dynamics[29]. We represent each individual in the population by a node in a complex network and draw links between nodes to indicate their embedding in a social group of others (see Methods section below for details). This relates directly to the idea of a *sociomatrix* that accounts for the stronger influence that individuals to which one forms a social bond have on one's behaviour[35]. In addition to the original formulation of this network cascade model[29] and in agreement with the consideration put forward above we assume that P randomly distributed nodes form the potentially active population. Being potentially active subsumes all norms, preferences and attitudes that cause an individual to show acceptance for a considered type of behaviour. Among the P potentially active nodes we assume that $A \leq P$ randomly distributed nodes are certainly active. In each time step each of the remaining $C = P - A$ contingent nodes i becomes active if more than a share $\rho \in [0, 1]$ of its immediate neighbors is already active. We hence denote ρ the *threshold fraction* of an individual. The resulting number or active nodes at time t is again denoted as $R(t)$. Setting a common value of ρ represents the most narrow distribution of actual threshold fractions that determine whether one joins into an action given that one generally supports that action at all.

We simulate cascades of nodes becoming active for two different shares of potentially active nodes $p = P/N = 0.56$ (Fig. 2a) and $p = 1$ (Fig. 2b), as well as for different threshold fractions $\rho \in \{0.2, 0.5, 0.8\}$. Figure 2 shows the final share of acting nodes $r^* = R^*/N$ after the cascade stops for increasing shares of certainly acting nodes $a = A/N \leq p$. For $p = 0.56$ (i.e., a low share of potentially acting nodes) only small threshold fractions ($\rho = 0.2$) allow for a large-scale cascade such that $r^* \to p$ for values of $a \gtrapprox 0.05$ (Fig. 2a). In contrast, for values of $a \lessapprox 0.05$ no cascade is observed and, hence, $r^* \lessapprox a$. Larger threshold fractions (i.e., $\rho = 0.5$ or $\rho = 0.8$) hinder the emergence of a cascade such that $r^* \lessapprox a$ for all choices of a (Fig. 2a). For $p = 1$, cascades are also observed at a larger threshold fraction of $\rho = 0.5$ but are still suppressed for $\rho = 0.8$ (Fig. 2b). Furthermore, the

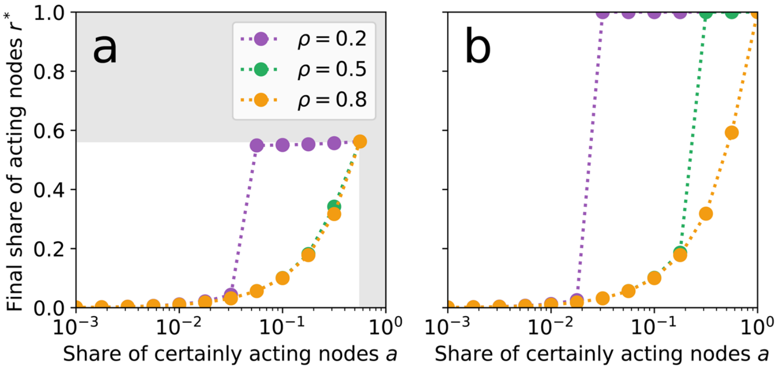


Figure 2. The final share of acting nodes r^* in the microscopic network simulation for given shares of certainly acting nodes a. (**a**) With only around half the population being potentially active (i.e, $p = P/N \approx 0.56$) only a low threshold fraction ($\rho = 0.2$, purple) causes large shares of the contingent nodes to act. Grey areas indicate values of r^* and a that would exceed p. (**b**) If every node in the network is potentially active ($p = 1$), also an intermediate threshold fraction ($\rho = 0.5$, green) suffices to cause the entire population to act. In comparison with (**a**) one also observes that the transition observed for $\rho = 0.2$ occurs already for smaller choices of a. For a large threshold fraction ($\rho = 0.8$, yellow) no abrupt transition appears such that $r^* \lessapprox a$ for all considered choices of a and p.

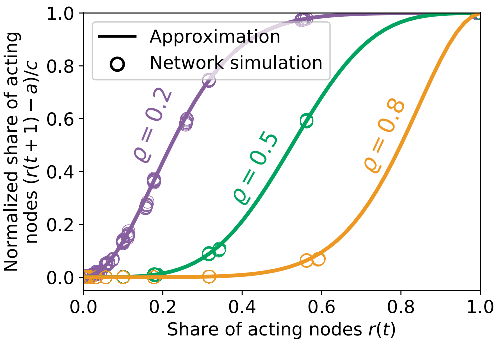


Figure 3. Emergent threshold distribution measured from the microscopic network simulations and the analytical approximation. For the network simulations only those points where the system is close to equilibrium, i.e. $t \in \{0, t_{max} - 1\}$, are shown. For all shown choices of threshold fractions ρ, the approximation matches well with the network simulations.

required share of certainly acting nodes a at which the system *tips* from a state with no cascades to a state with a global cascade decreases slightly with increasing p (compare Fig. 2a,b). Note that specifically the role of the remaining $N - P$ certainly inactive nodes has been studied under the term '*immune* nodes' in an earlier study of spreading dynamics on networks[43]. However, in contrast to our results presented above the underlying model in this previous work[43] assumed the share of certainly active nodes a to increase over time at a constant rate, thus yielding convergence to a globally stable fixed point $r^* = p$ for all initial choices of a. Hence, the major purpose of the *immune* nodes in this earlier work was to moderate the rate of convergence to that global fixed point.

To estimate an emergent threshold distribution as required for the Granovetter-type threshold model we now evaluate $r(t) = R(t)/N$ against $(r(t+1) - a)/c$ (with $c = C/N = (P - A)/N$) from the network simulations. Figure 3 shows the results if the network cascade is close to equilibrium, i.e., for $t = 0$ or $t = t_{max} - 1$, where t_{max} is the time at which the cascade stops. We observe the formerly postulated broad distribution of emergent thresholds as a result of the microscopic interactions at narrowly distributed threshold fractions $\rho \in \{0.2, 0.5, 0.8\}$ given a generally positive (P nodes) or negative ($N - P$ nodes) attitude towards the considered behavior. This implies that individuals with a high emergent threshold may not necessarily be more reluctant to join into an action, it could simply mean that they are located at a more peripheral position in the network.

By approximating the number of active, a_i, and inactive neighbors, b_i, of a node i as coming from a common multinomial distribution that only depends on the number of neighbors $k_i = a_i + b_i$ and the overall share of active nodes $r(t)$, we derive an analytical approximation of the emergent threshold distribution F (note that for brevity we omit the dependence of $r(t)$ on t) as

Reproduced from: M. Wiedermann et al., A network-based microfoundation of Granovetter's threshold model for social tipping, Sci. Rep., vol. 10, no. 1, pp. 1-10, 2020, doi: 10.1038/s41598-020-67102-6. Published under Creative Commons Attribution License 4.0 (CC BY).

$$F(r) = 1 - \exp(-K) \sum_{b_i=0}^{\infty} \frac{(K - Kr)^{b_i}}{b_i!} \sum_{a_i=0}^{\left[\frac{\rho b_i}{1-\rho}\right]} \frac{(Kr)^{a_i}}{a_i!}. \tag{3}$$

here, $K = \sum_i k_i / N$ denotes the average degree (i.e., number of neighbors) of nodes in the network (see Methods section below and the Supplementary Information for a full derivation of Eq. (3)). Note that the second factor in Eq. (3) can be further approximated by an incomplete gamma function. We find that (close to equilibrium) Eq. (3) aligns very well with the network simulations for small ($\rho = 0.2$), medium ($\rho = 0.5$) and large ($\rho = 0.8$) fractional thresholds (Fig. 3) and thus complements previously proposed approximations that primarily held for small to medium values[42]. For the transient phase the approximation still estimates the emergent thresholds well for small and large choices of ρ but decreases in quality for intermediate values (see Supplementary Information). This is mainly caused by the clustering of active and inactive nodes. An extension of the above approximation that accounts for such factors, e.g., via pair approximations[54,55] or moment generating functions[29], is beyond the scope of this work and remains as a subject for future research. In summary, Eq. (3) gives a good estimation of an emergent macroscopic distribution that fulfills the initially postulated broad shape[35] while emerging from a subsumed set of preferences as well as a single common threshold fraction ρ. In addition, using a single distribution F has the advantage of being independent of the share of certainly and potentially acting nodes. As such it only needs to be estimated once while changing preferences (i.e., varying A and P) are incorporated into shifting the diagonal line that is used to estimate the fixed points (see again Fig. 1c).

Comprehensive analysis and social tipping. From the approximate emergent threshold distribution F in Eq. (3) we estimate the fixed points r^* of the refined threshold model for different choices of a, p (or $c = p - a$), and ρ by solving $(r - a)/c = F(r)$ (i.e., intersecting the diagonal line with F). We either identify two stable and one unstable fixed points, or one globally stable fixed point r^*. Figure 4a shows the value of the smallest stable fixed point $\min(r^*)$. We find a sharp increase in its value for certain values of $0.15 \lessapprox a \lessapprox 0.22$ and $p \gtrapprox 0.5$ hinting at a saddle-node bifurcation. Figure 4b,c show that saddle-node bifurcation at varying values of a and p, respectively. As the saddle-node bifurcation, and correspondingly also hysteresis, emerges in both parameters, the model consequently displays a cusp bifurcation as well (see black circle in Fig. 4a). For fixed values of a or p below the cusp-point the final share of acting individuals r^* thus varies only smoothly with the respective other free parameter (red lines in Fig. 4b,c). In contrast, fixing either a or p to values above the cusp-point can cause the system to rapidly shift from a stable state with low r^* to a stable state with high r^* (and vice versa) as the corresponding bifurcation point in the remaining free parameter is crossed (black lines in Fig. 4b,c). Notably, the model shows hysteresis also within a band of possible threshold fractions, Fig. 4d.

In summary, our model conceptually shows what has formerly been termed *social tipping*, i.e., a process where, for a given population, a small change in the size of a dedicated minority can have a large effect[19,21,56]. In our specific case, for a given value of a or p a small change in the respective other parameter suffices to largely increase (or decrease) the share of finally acting individuals r^*. Complementing recent theoretical and numerical studies of spreading processes on networks that either varied the size of the initiating minority[42] or the so-called immune group of inactive nodes[43] our model shows a bistable regime that is necessary for the emergence of hysteresis. This implies that once the system has tipped it sustains its state of high (low) shares of acting individuals r^* even if a or p were to be reduced (increased) again. By incorporating both, initiating and immune groups, our model additionally gives rise to a previously undetected cusp bifurcation as well.

Remarkably, the critical size of the dedicated minority at which the system undergoes a fold bifurcation (Fig. 4a,b) has recently been empirically estimated to lie in the range $0.21 \lessapprox a \lessapprox 0.25$ which is consistent with the results of our model[19]. Moreover, critical minority group sizes of around 20 percent have also been discussed with respect to the Pareto principle[57] which has recently been reframed as *the law of the vital few* to discuss matters of sustainability transformations and social tipping[58].

Discussion

We have proposed a refined version of the original Granovetter threshold model[35] that addresses a set of issues that, so far, have hindered its application as a conceptual modeling tool. Specifically, we propose to divide the considered population of size N into three classes (certainly, potentially, and certainly not acting individuals) of different sizes $A \leq P, P \leq N$, and $N - P$. In addition, we propose a threshold distribution that emerges from microscopic interactions between individuals on a social network. This distribution solely depends on the average connectivity K of individuals and a common threshold fraction ρ to join into an action given that their individual preferences and attitudes are already favourable with respect to that action. The four parameters of our refined model are of intuitive nature and allow for a systematic evaluation of its dynamics in terms of a bifurcation analysis (except for K which only needs to be chosen sufficiently larger than zero, i.e., $K \gg 0$, see Supplementary Information for details). As in the original threshold model, an estimation of the fixed points can be obtained by (graphically) intersecting the diagonal line defined by a and p with the emergent threshold distribution F. The three crucial parameters a, p, and ρ all cause a saddle-node bifurcation which is a prototypical mechanism behind tipping points in many other systems, such as in ecology[59,60] or the climate system[61,62], as well. It thus makes the model a promising tool to study the emerging field of social tipping[19,21,56] where *little things can make a big difference*[63] and minority groups can trigger large shares of a population to engage in collective action.

Our revised model describes multiple forms of collective behaviors, including social movements and crowd-like behaviors. For both such behaviors, norms are directly called upon to structure individual likelihood

Reproduced from: M. Wiedermann et al., A network-based microfoundation of Granovetter's threshold model for social tipping, Sci. Rep., vol. 10, no. 1, pp. 1-10, 2020, doi: 10.1038/s41598-020-67102-6. Published under Creative Commons Attribution License 4.0 (CC BY).

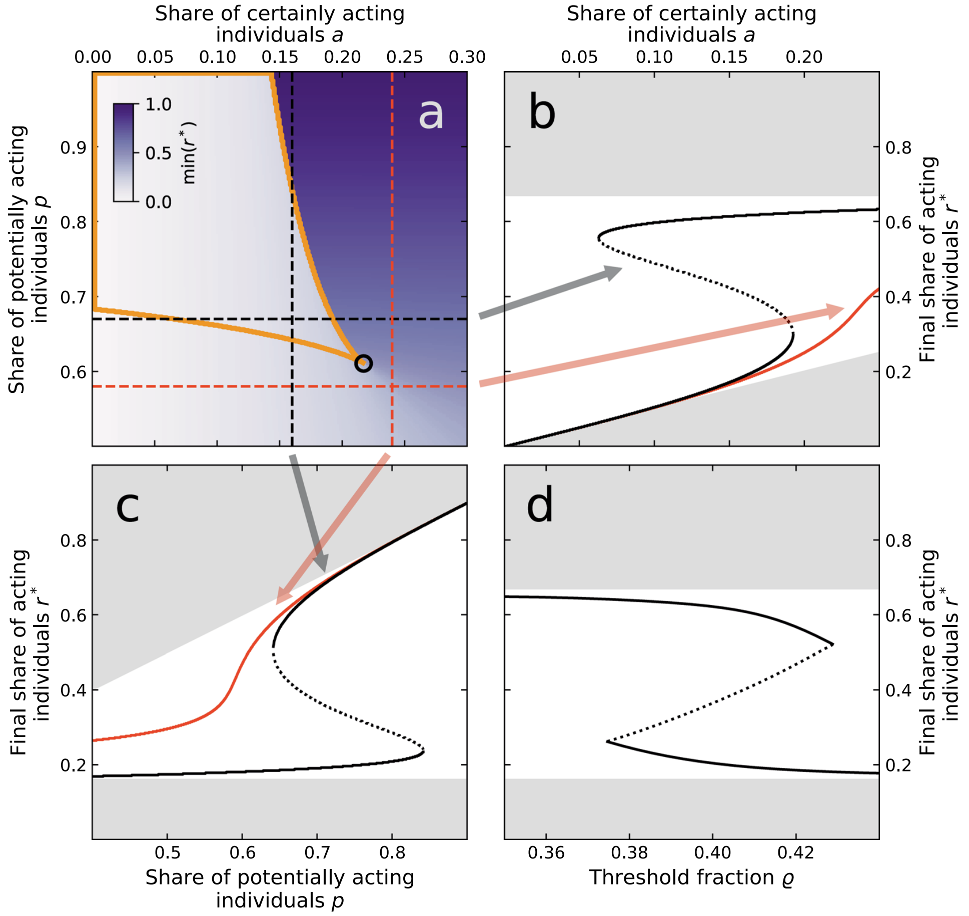


Figure 4. Bifurcation analysis and hysteresis of the refined Granovetter model with an emergent threshold distribution as given by the analytical approximation. (**a**) Smallest stable fixed point min(r^*) for different shares of certainly acting a and potentially acting individuals p. The black circle denotes a cusp-bifurcation. Black dashed horizontal/vertical lines correspond to the diagrams in (**b**,**c**) that show a saddle-node bifurcation. For (**b**–**d**), solid (dotted) lines indicate stable (unstable) fixed points r^*. Grey shading indicates those areas where $r^* \notin [a, p]$ and that can thus not be reached. The yellow circled area in (**a**) indicates the bistable regime. Red dashed horizontal/vertical lines in (**a**) correspond to values of p and a at which no bifurcation is observed and thus r^* varies smoothly in (**b**,**c**). (**d**) Shows the bifurcation diagram in the threshold fraction ρ. Fixed parameters are: $a = 0.16$ for (**c**) ($a = 0.24$ for the red curve) and (**d**), $p = 0.67$ for (**b**) ($p = 0.58$ for the red curve) and (**d**), and $\rho = 0.4$ for (**a**–**c**).

to engage in actions while also observing the actions of others around them. Importantly, there are differences in the speed of the process. For crowds the observation of social members is made relatively quickly, as are the decisions to participate in the actions. In contrast, these processes can be much slower for social movements. For both cases, however, we identify three time scales that are underlying our refined threshold model. We assume that the microscopic threshold fractions change at the slowest time scale (usually years to decades), as these are attributed to the unique identity of an individual (which may be less prone to sudden external shocks). In contrast, the classification into *certainly* or *contingently* active individuals varies on intermediate time scales (months to years) as changes in the environment (such as financial shocks or the exposition to increasing extreme weather events) are beyond an individual's own agency and can trigger sudden changes in attitudes[64]. The social dynamics modelled here, i.e., the observation of others and the joining into an action, are happening on the fastest time scale (days to months) as frequent social interactions are common among members of any given society.

Most parameters of the refined model may be readily measurable in a variety of applications. Attitudes that determine p could be estimated from surveys or existing panel data. The share of certainly acting individuals a could be given by those in the population that inevitably need to act, e.g., migrate as a consequence of climate change impacts[65,66]. For the average degree K it may often suffice to set it to a reasonable number, e.g., Dunbar's number that suggests a cognitive limit to the number of people with whom an individual can maintain a persistent social relationship[67] (see Supplementary Information for details). The threshold fraction ρ could then either

Reproduced from: M. Wiedermann et al., A network-based microfoundation of Granovetter's threshold model for social tipping, Sci. Rep., vol. 10, no. 1, pp. 1-10, 2020, doi: 10.1038/s41598-020-67102-6. Published under Creative Commons Attribution License 4.0 (CC BY).

remain as a free parameter of the model or be set to fixed intuitive points such as 50% (majority rule) or 20% (Pareto principle[57,58]). Furthermore, the model also allows for changes in its parameters over time, such that r^* can be estimated as a time-dependent variable, possibly causing the system to tip back and forth between its two possible stable states. In that sense the respective parameters can be incorporated into the system's internal dynamics as slowly changing variables.

Future work should concentrate on collecting data for the different parameters and then consequently test and calibrate the model against historical test cases. One specific challenge that lies within such an endeavor is the estimation of appropriate (relative) time scales at which the parameters and the internal variables change. In addition, appropriate early-warning indicators[62,68,69] should be applied to study the existence of precursory signals for the transgression of a social tipping point, i.e., bifurcation, in our model. Some of these indicators would require a further extension of the model such that individuals may also spontaneously become active with a low probability even if their threshold fraction is not transgressed (or vice versa). We further acknowledge that up to now a proposal for an emergent threshold distribution has only been derived analytically for the case of an Erdős—Rényi random network[70]. While this lays good groundwork, the threshold distribution should also be explored for topologies (such as *scale-free*[71] and *small-world* networks[72]) that more closely mimic those of real-world social systems. Hence, even though our proposed approximation of the emergent threshold distribution holds well if the system is well-mixed and close to a fixed point, more elaborate methods, e.g., pair approximations[55] and moment generating function approaches[29], should be used to predict the model's dynamics for more general network topologies and during transient phases as well. Ultimately, the model should be applied as a conceptual modeling tool, e.g., to make qualitative statements on the possibility for social tipping with respect to issues of global change or sustainability transformations[12,73,74] under different scenarios.

Methods

Network cascade model. For the microscopic network simulation we consider an Erdős—Rényi random network[70] with $N = 100\,000$ nodes and a linking probability of $\ell = 9 \cdot 10^{-5}$ resulting in an average degree of $K = 10$. We vary the number of certainly acting nodes A logarithmically between 1 and N and the number of potentially acting nodes logarithmically between A and N. For each setting of A and P (and fixed values of the threshold fraction ρ as given in Fig. 2) we create an ensemble of $n = 100$ networks and randomly assign P out of the N nodes as potentially active. Out of those P nodes we then randomly assign A certainly acting nodes. The model then runs in discrete time steps t. In each time step, every potentially active, yet inactive, node i becomes active if its share of active neighbors exceeds the threshold fraction ρ. All nodes update their status synchronously at each time step. The simulation stops if the number of newly activated nodes at time t equals zero, i.e., if $R(t-1) = R(t)$. Note that our model is based on previous works that implemented a simpler version of a cascade model that did not account for a distinction in potentially active and certainly inactive nodes[29].

Approximation of the emergent threshold distribution. The approximate emergent threshold distribution F in Eq. (3) is derived by assuming that for each individual i the number of active a_i and inactive neighbors b_i are distributed according to a common multinomial distribution, giving

$$F(R) = \sum_{\substack{a_i > \rho(a_i+b_i)\\ a_i \le R\\ b_i \ge 0\\ b_i \le P'}} \binom{R}{a_i}\binom{P'}{b_i} \ell^{a_i}\ell^{b_i}(1-\ell)^{R-a_i}(1-\ell)^{P'-b_i}. \tag{4}$$

$P' = N - 1 - R$ denotes the number of inactive individuals that are not the considered i, as one's own level of activity is not accounted for. ℓ is the linking probability of the Erdős—Rényi network. Equation (3) follows from Eq. (4) by setting $R = \lfloor rN \rfloor$, substituting the binomial distributions by two Poisson distributions with expectation values $\lambda_a = Kr$ and $\lambda_b = K - Kr$ and assuming that $N \gg K$. A step-by-step derivation of Eq. (3) is given in the Supplementary Information.

Received: 27 March 2020; Accepted: 2 June 2020;
Published online: 08 July 2020

References

1. Snow, D. & Oliver, P. Social Movements and Collective Behavior: Social Psychological Dimensions and Considerations. In Cook, K., Fine, G. & House, J. (eds.) *Sociological Perspectives on Social Psychology*, 571–599 (Allyn and Bacon, Needham Heights, MA, 1995).
2. Park, R. *The Crowd and the Public*. (University of Chicago Press, Chicago, 1904).
3. Blumer, H. Collective Behavior. In Park, R. (ed.) *Principles of Sociology*, 219–288, 2nd edn. (Barnes and Noble, New York, 1939).
4. Lofland, J. *Protest: Studies of Collective Behavior and Social Movements*. (Transaction Books, New Brunswick, N.J., 1985).
5. McPhail, C. *The Myth of the Madding Crowd*. (Routledge, New York, 1991).
6. Diani, M. The Concept of Social Movement. *The Sociological Review* **40**, 1–25 (1992).
7. Snow, D., Soule, S. & Kriesi, H. Mapping the Terrain. In Snow, D., Soule, S. & Kriesi, H. (eds.) *The Blackwell Companion to Social Movements*, 3–16 (Blackwell Publishing, Malden, MA, 2004).
8. McPhail, C. The Crowd and Collective Behavior: Bringing Symbolic Interaction Back In. *Symbolic Interaction* **29**, 433–464 (2006).
9. Nyborg, K. *et al.* Social norms as solutions. *Science* **354**, 42–43 (2016).
10. McPhail, C. Blumer's Theory of Collective Behavior: The Development of a Non-Symbolic Interaction Explanation. *The Sociological Quarterly* **30**, 401–423 (1989).
11. Hagedorn, G. *et al.* Concerns of young protesters are justified. *Science* **364**, 139–140 (2019).
12. Farmer, J. D. *et al.* Sensitive intervention points in the post-carbon transition. *Science* **364**, 132–134 (2019).

13. Moser, S. C. & Dilling, L. Toward the social tipping point: Creating a climate for change. *Creating a climate for change: Communicating climate change and facilitating social change* 491–516 (2007).
14. Otto, I. M. *et al*. Social tipping dynamics for stabilizing Earth's climate by 2050. *Proceedings of the National Academy of Sciences* **117**(5), 2354–2365 (2020).
15. Mønsted, B., Sapieżyński, P., Ferrara, E. & Lehmann, S. Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PLoS One* **12**, e0184148 (2017).
16. Centola, D. The Spread of Behavior in an Online Social Network Experiment. *Science* **329**, 1194–1197 (2010).
17. Márton, K., Gerardo, I., Kimmo, K. & János, K. Complex contagion process in spreading of online innovation. *Journal of The Royal Society Interface* **11**, 20140694 (2014).
18. Christakis, N. A. & Fowler, J. H. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in Medicine* **32**, 556–577 (2013).
19. Centola, D., Becker, J., Brackbill, D. & Baronchelli, A. Experimental evidence for tipping points in social convention. *Science* **360**, 1116–1119 (2018).
20. Milkoreit, M. *et al*. Defining tipping points for social-ecological systems scholarship—an interdisciplinary literature review. *Environmental Research Letters* **13**, 033005 (2018).
21. Bentley, R. A. *et al*. Social tipping points and Earth systems dynamics. *Frontiers in Environmental Science* **2** (2014).
22. House Thomas. Modelling behavioural contagion. *Journal of The Royal Society Interface* **8**, 909–912 (2011).
23. Guilbeault, D., Becker, J. & Centola, D. Complex Contagions: A Decade in Review. In Lehmann, S. & Ahn, Y.-Y. (eds.) *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*, Computational Social Sciences, 3–25 (Springer International Publishing, Cham, 2018).
24. Vespignani, A. Modelling dynamical processes in complex socio-technical systems. *Nature Physics* **8**, 32–39 (2012).
25. Melnik, S., Ward, J. A., Gleeson, J. P. & Porter, M. A. Multi-stage complex contagions. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **23**, 013124 (2013).
26. Watts, D. J. & Dodds, P. S. Influentials, Networks, and Public Opinion Formation. *Journal of Consumer Research* **34**, 441–458 (2007).
27. Dodds, P. S. & Watts, D. J. Universal Behavior in a Generalized Model of Contagion. *Physical Review Letters* **92**, 218701 (2004).
28. Dodds, P. & Watts, D. A generalized model of social and biological contagion. *Journal of Theoretical Biology* **232**, 587–604 (2005).
29. Watts, D. J. A Simple Model of Global Cascades on Random Networks. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5766–5771 (2002).
30. Hołyst, J. A., Kacperski, K. & Schweitzer, F. Social impact models of opinion dynamics. In *Annual Reviews of Computational Physics IX*, 253–273 (2001).
31. Hegselmann, R. & Krause, U. Opinion Dynamics Driven by Various Ways of Averaging. *Computational Economics* **25**, 381–405 (2005).
32. Schleussner, C.-F., Donges, J. F., Engemann, D. A. & Levermann, A. Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure. *Scientific Reports* **6**, 30790 (2016).
33. Aoki, I. A Simulation Study on the Schooling Mechanism in Fish. *Nippon Suisan Gakkaishi* **48**, 1081–1088 (1982).
34. Couzin, I. D., Krause, J., James, R., Ruxton, G. D. & Franks, N. R. Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology* **218**, 1–11 (2002).
35. Granovetter, M. Threshold Models of Collective Behavior. *American Journal of Sociology* **83**, 1420–1443 (1978).
36. Kaempfer, W. H. & Lowenberg, A. D. A Threshold Model of Electoral Policy and Voter Turnout. *Rationality and Society* **5**, 107–126 (1993).
37. Zeppini, P., Frenken, K. & Kupers, R. Thresholds models of technological transitions. *Environmental Innovation and Societal Transitions* **11**, 54–70 (2014).
38. Hunter, L. M. Migration and Environmental Hazards. *Population and Environment* **26**, 273–302 (2005).
39. Lohmann, S. The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989–91. *World Politics* **47**, 42–101 (1994).
40. McCarthy, J. D. & Zald, M. N. Resource mobilization and social movements: A partial theory. *American Journal of Sociology* **82**, 1212–1241 (1977).
41. Jenkins, J. C. Resource mobilization theory and the study of social movements. *Annual review of sociology* **9**, 527–553 (1983).
42. Singh, P., Sreenivasan, S., Szymanski, B. K. & Korniss, G. Threshold-limited spreading in social networks with multiple initiators. *Scientific Reports* **3**, 1–7 (2013).
43. Karsai, M., Iñiguez, G., Kikas, R., Kaski, K. & Kertész, J. Local cascades induced global contagion: How heterogeneous thresholds, exogenous effects, and unconcerned behaviour govern online adoption spreading. *Scientific Reports* **6**, 1–10 (2016).
44. Centola, D. & Baronchelli, A. The spontaneous emergence of conventions: An experimental study of cultural evolution. *Proceedings of the National Academy of Sciences* **112**, 1989–1994 (2015).
45. Garrod, S. & Doherty, G. Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition* **53**, 181–215 (1994).
46. Strang, D. & Soule, S. A. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual review of sociology* **24**, 265–290 (1998).
47. DiMaggio, P. J. & Powell, W. W. The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American sociological review* 147–160 (1983).
48. Corning, A. F. & Myers, D. J. Individual orientation toward engagement in social action. *Political Psychology* **23**, 703–729 (2002).
49. Paulsen, R. Education, social class, and participation in collective action. *Sociology of Education* **64**, 96–110 (1991).
50. Lim, C. Social Networks and Political Participation: How Do Networks Matter? *Social Forces* **87**, 961–982 (2008).
51. Schelling, T. C. Hockey Helmets, Concealed Weapons, and Daylight Saving: A Study of Binary Choices With Externalities. *Journal of Conflict Resolution* **17**, 381–428 (1973).
52. Schelling, T. C. Dynamic models of segregation. *The Journal of Mathematical Sociology* **1**, 143–186 (1971).
53. Newman, M. *Networks: An Introduction* (OUP Oxford, 2010).
54. Wiedermann, M., Donges, J. F., Heitzig, J., Lucht, W. & Kurths, J. Macroscopic description of complex adaptive networks coevolving with dynamic node states. *Physical Review E* **91**, 052801 (2015).
55. Gleeson, J. P. Binary-State Dynamics on Complex Networks: Pair Approximation and Beyond. *Physical Review X* **3**, 021004 (2013).
56. Pruitt Jonathan, N. *et al*. Social tipping points in animal societies. *Proceedings of the Royal Society B: Biological Sciences* **285**, 20181282 (2018).
57. Pareto, V. *Manual of political economy*. (A. M. Kelley, New York, 1971).
58. Schellnhuber, H. J., Rahmstorf, S. & Winkelmann, R. Why the right climate target was agreed in Paris. *Nature Climate Change* **6**, 649–653 (2016).
59. Beisner, B. E., Haydon, D. T. & Cuddington, K. Alternative stable states in ecology. *Frontiers in Ecology and the Environment* **1**, 376–382 (2003).
60. Dai, L., Vorselen, D., Korolev, K. S. & Gore, J. Generic Indicators for Loss of Resilience Before a Tipping Point Leading to Population Collapse. *Science* **336**, 1175–1177 (2012).
61. Lenton, T. M., Livina, V. N., Dakos, V. & Scheffer, M. Climate bifurcation during the last deglaciation? *Climate of the Past* **8**, 1127–1139 (2012).

Reproduced from: M. Wiedermann et al., A network-based microfoundation of Granovetter's threshold model for social tipping, Sci. Rep., vol. 10, no. 1, pp. 1-10, 2020, doi: 10.1038/s41598-020-67102-6. Published under Creative Commons Attribution License 4.0 (CC BY).

62. Thompson, J. M. T. & Sieber, J. Predicting climate tipping as a noisy bifurcation: a review. *International Journal of Bifurcation and Chaos* **21**, 399–423 (2011).
63. Gladwell, M. *The tipping point: How little things can make a big difference* (Little, Brown, 2006).
64. Ricke, K. L. & Caldeira, K. Natural climate variability and future climate policy. *Nature Climate Change* **4**, 333–338 (2014).
65. Black, R., Bennett, S. R. G., Thomas, S. M. & Beddington, J. R. Climate change: Migration as adaptation. *Nature* **478**, 447–449 (2011).
66. McLeman, R. & Smit, B. Migration as an Adaptation to Climate Change. *Climatic Change* **76**, 31–53 (2006).
67. Dunbar, R. I. M. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences* **16**, 681–694 (1993).
68. Scheffer, M. *et al.* Early-warning signals for critical transitions. *Nature* **461**, 53–59 (2009).
69. Jiang, J. *et al.* Predicting tipping points in mutualistic networks through dimension reduction. *Proceedings of the National Academy of Sciences* **115**, E639–E647 (2018).
70. Erdős, P. & Rényi, A. On the Evolution of Random Graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 17–61 (1960).
71. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286**, 509–512 (1999).
72. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393**, 440–442 (1998).
73. Westley, F. *et al.* Tipping Toward Sustainability: Emerging Pathways of Transformation. *Ambio* **40**, 762–780 (2011).
74. David Tàbara, J. *et al.* Positive tipping points in a rapidly warming world. *Current Opinion in Environmental Sustainability* **31**, 120–129 (2018).

Acknowledgements

This work was developed in the context of the COPAN collaboration at the Potsdam Institute for Climate Impact Research (PIK). M.W. and K.S. are supported by the Leibniz Association (project DOMINOES). J.F.D. is grateful for financial support by the Earth League's EarthDoc program and the European Research Council advanced grant project ERA (Earth Resilience in the Anthropocene). The authors gratefully acknowledge the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the Land Brandenburg for providing resources on the high-performance computer system at PIK.

Author contributions

All authors designed the study. M.W. performed the numerical simulations and analysed the data. M.W. and J.H. derived the analytical approximation. M.W. and E.K.S. drafted the manuscript. All authors substantively revised the work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-67102-6.

Correspondence and requests for materials should be addressed to M.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Received: 17 February 2016
Accepted: 11 July 2016
Published: 11 August 2016

Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure

Carl-Friedrich Schleussner[1,2,*], Jonathan F. Donges[2,3,*], Denis A. Engemann[4,5,*] & Anders Levermann[2,6,7]

Large-scale transitions in societies are associated with both individual behavioural change and restructuring of the social network. These two factors have often been considered independently, yet recent advances in social network research challenge this view. Here we show that common features of societal marginalization and clustering emerge naturally during transitions in a co-evolutionary adaptive network model. This is achieved by explicitly considering the interplay between individual interaction and a dynamic network structure in behavioural selection. We exemplify this mechanism by simulating how smoking behaviour and the network structure get reconfigured by changing social norms. Our results are consistent with empirical findings: The prevalence of smoking was reduced, remaining smokers were preferentially connected among each other and formed increasingly marginalized clusters. We propose that self-amplifying feedbacks between individual behaviour and dynamic restructuring of the network are main drivers of the transition. This generative mechanism for co-evolution of individual behaviour and social network structure may apply to a wide range of examples beyond smoking.

Behaviour is shaped by interactions between the individual and its environment[1]. As a result of evolutionary pressures emanating from intensified group lifestyles, humans acquired a diverse and specialised social behavioural repertoire[2,3]. The human cognitive capacity for enduring collaborative social interaction has been extensively investigated in various disciplines related to the field of cognitive sciences (for an overview cf. refs 4–6). Theories about how behaviours are shaped by social and individual factors have a longstanding tradition in psychology and social sciences[7,8]. Examples of quantitative models include dynamic models of segregation in urban neighbourhoods[9], models of cultural dissemination[10] and a wealth of literature aiming at the inclusion of social decision making into economic theory[11,12]. An overview on mathematical approaches to social dynamics is provided in ref. 13. Importantly, social relations can be represented as graphs, which renders them accessible to network theoretical analysis[14]. It has been shown that the overall structure of connections between individuals in social networks and face-to-face interactions between individuals systematically affects a wide range of social and individual characteristics, such as happiness, divorce rates, smoking and obesity[15–19]. We refer to this effect as behaviour selection emphasising an evolutionary process rather than mere individual decision-making.

In this context, networks statistics enable more targeted characterisations of social dynamics. For example, social distance modulates similarity and behavioural synchrony between individuals. It is commonly measured using the shortest path length between two individuals in a social network and has been shown to preferentially shape individual behaviour up to a distance of three social ties[19]. Likewise, the contents of social interactions

[1]Climate Analytics, Berlin, Germany. [2]Potsdam Institute for Climate Impact Research, Potsdam, Germany. [3]Stockholm Resilience Centre, Stockholm University, Stockholm, Sweden. [4]Cognitive Neuroimaging Unit, CEA DRF/I2BM, INSERM, Université Paris-Sud, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France. [5]Neuropsychology & Neuroimaging Team, INSERM UMRS 975, ICM, Paris, France. [6]Lamont-Doherty Earth Observatory, Columbia University, New York, USA. [7]Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany. [*]These authors contributed equally to this work. Correspondence and requests for materials should be addressed to C.-F.S. (email: schleussner@pik-potsdam.de) or J.F.D. (email: donges@pik-potsdam.de) or D.A.E. (email: denis.engemann@gmail.com)

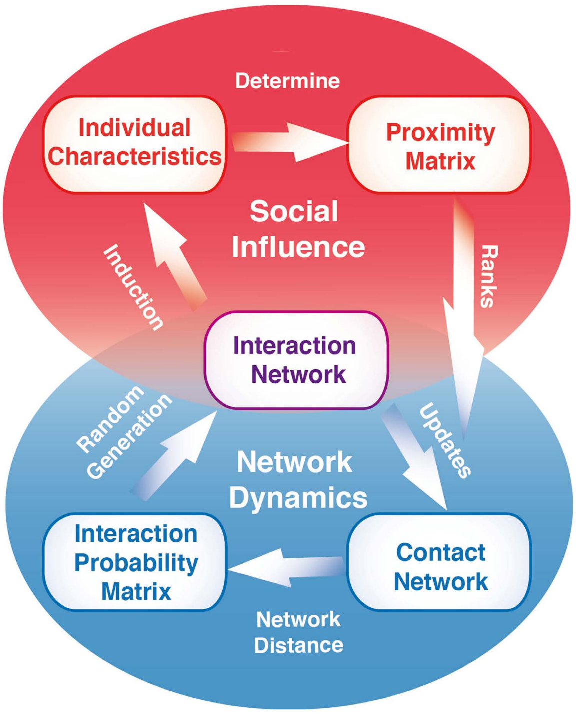


Figure 1. Adaptive network model of behaviour selection. For a group of individuals, the proposed model predicts selection of behaviour as a function of two factors: local interaction between individuals and the global structure of their social connections. The *proximity matrix* describes how similar a given pair of individuals is based on their *individual characteristics* such as smoking behaviour. Assuming restricted resources, agents maintain a limited number of social contacts. In the proposed model, individuals only keep their most proximate contacts. Based on the *proximity matrix*, it can be determined which individuals are current neighbours in the *contact network*. The distance between nodes in this network is then used to compute the *interaction probability matrix* for stochastically generating current interactions between a given pair of individuals. Importantly, this *interaction network* exerts feedback on the individual behaviour and thus closes the co-evolutionary loop: The probability of changing the smoking behaviour is modelled as a function of the individual smoking disposition and the dominance of smoking behaviour in the local neighbourhood of the interaction network. Note that only individuals who have actually interacted can establish a tie in the contact network in the next time step.

between individuals depend on their relationship, i.e., perceived friendship status, which suggests nonlinear interdependencies between network dynamics and social interaction[20]. Opinion formation and imitation have been advocated as candidate mechanisms of behaviour induction[21,22]. This puts emphasis on cognitive processes and biases for selective spread of behaviours in social networks[23,24].

These findings motivate process-based techniques for modelling dynamical social networks highlighting time-varying aspects of social connections[25]. One such approach is represented by adaptive networks that model the temporal co-evolution of network structure and dynamic node states[26–28]. The most commonly modelled social processes include imitation, collaboration dynamics and social tie formation as a function of antipathy and sympathy (sometimes referred to as homophily and heterophily, respectively) between agents. Adaptive network models have then been used to investigate complex phenomena in social networks such as phase transitions and tipping points in opinion formation on single-[29] and multi-layer[30] networks, epidemic spreading[31], swarm behaviour[32], friendship structure in social media networks[33], sustainable use of renewable resources[34] and coalition formation[35]. Opinion formation has been intensively investigated using the adaptive voter model[29], its generalisations and related models[26–28] with a focus on consensus formation[36], opinion diversity[37] and network fragmentation[38].

In adaptive network models, the selection of update rules critically determines the co-evolutionary dynamics of node states and network structure. When considering real-world social systems, social connections are very unlikely to be established randomly and in disregard of the underlying network structure, but rather are the outcome of agents' interactions in a complex network[39–41]. At the same time, interaction between agents along social ties is a key process to induce individual behavioural change[42]. Therefore, choosing update rules such that they explicitly take into account peculiarities of micro-scale social interactions seems promising for obtaining more

		Behaviour update by social interaction	
		Yes	**No**
Feedback into contact network	Yes	*coupled, mean-field*	*network*
	No	*interaction*	—

Table 1. Partial and alternative models of behaviour selection studied in this work.

realistic models. Such an interaction-resolved adaptive network approach would then also allow to study the co-evolution of micro-scale social influence and large-scale network structures.

More empirical findings have become available that explicitly describe social network structures. In the work presented here, we will focus on a study on smoking habits by Christakis and Fowler[17]. Based on a detailed long-term survey, they analysed smoking habits of 12,067 inhabitants of a small town in the US between 1971 and 2003 while concomitantly tracking their social relationship structure, i.e., mutual assessment of friendship status. Their analysis revealed that over that time period, the prevalence of smoking declined from about 50% to about 10%. At the same time, the structure of social connections changed almost selectively for the remaining smokers. Their average eigenvector centrality, a measure of how much a node is in the "centre" of its social network, significantly declined. At the same time, the probability of an individual being a smoker conditional on the prevalence of smoking in its neighbourhood (referred to as *conditional probability* below, see Methods) increased up to the level of third degree contacts (contacts of contacts of contacts). In other words, individuals who did not adapt to the decreasing societal support for smoking preferentially interacted with similar individuals, forming subgroups or clusters of increasingly marginalized smokers.

An Adaptive Network Model of Behaviour Selection

In the following, we will introduce an interaction-resolved adaptive network approach that we evaluate in terms of its capacity to reproduce characteristics of empirically studied time-varying social networks. We first outline the general modelling framework that contains core conceptual ideas of our adaptive network model of behaviour selection presented in Fig. 1. In a next step, we describe specific modifications and additions made to model social dynamics of changing smoking behaviour to reproduce findings from the empirical reference case[17].

Complex systems, such as the human brain, social networks, or the backbone structure of the internet, typically implement functional hierarchies[43–46]. Although dynamics with multiple temporal hierarchies also apply to the emergence of complex macroscopic structure in social systems in which agents repeatedly interact over time[47–49], hierarchical social network dynamics have rarely been explicitly modelled[26,27,50,51]. Here we considered functional hierarchies as coupling between an interaction network with fast updates and contact network with slow updates that together shape individual characteristics as their states change over time with preferential formation of social ties. A schematic overview of the model and its components is depicted in Fig. 1 and a detailed formal description of our model is given in the methods description below.

The *contact network*'s structure is based on an overall similarity between individual's characteristics such as preferences, socio-economic status or genetic factors (cf. refs 52–55) that generate a social proximity between individuals. Here, contacts are understood as the number of other agents an individual may regularly interact with (a counterexample for this are entries in a Facebook contact list that only require a single interaction to be established). The total number of such contacts that can be maintained by a human individual is constrained by temporal and cognitive capacities[2]. General cognitive capacity and the number of contacts are both subject to individual differences[56,57]. We therefore restricted the maximum degree of social contact that an agent in the contact network is capable or willing to maintain by introducing an individual *degree preference* parameter that is normally distributed. An agent cannot maintain more contacts than prescribed by its degree preference, which implies that establishing new contacts (by a new edge in the contact network) may require disbanding old ones (deleting the edge in the contact network).

The *interaction network* provides the basis for establishing new contacts while at the same time also inducing change in the individual characteristics. It is generated stochastically at each time step based on the contact network. Reflecting empirical findings[19], the probability of interactions between two individuals in a given time step (represented by an edge in the interaction network) decreases with the shortest path distance between them in the contact network. The minimal interaction probability is a constant positive value, thereby allowing for unlikely, incidental meetings ("by chance") between distant or disconnected individuals.

In our model, tie formation in the contact network is constrained by the social proximity that may change as a result of the interaction[58]. To update the contact network, the social proximities to neighbours in the interaction network are compared with proximity values to contact network neighbours. Only the top ranking contacts are maintained, both in the contact and the interaction network up to the agent specific *degree preference*. Importantly, to establish a contact between two agents, each of them has to be included in the other's set of preferred contacts. By this requirement of reciprocity, previous contacts can be replaced actively, but also lost passively, reminiscent of forgetting. Such a process can be illustrated by an agent moving from city *A* to another city *B* in which she establishes new contacts and at the same time gradually forgets about her previous social network in *A*. In turn, also her previous contacts in *A* loose contact with her.

At the same time, the social influence dynamics play out on the interaction network. Individual characteristics such as behaviours are subject to peer-influence by direct neighbours in the interaction network as will be described below. The model design also allows to account for individual dispositions as node-dependent constraints on behaviour exogenous to the model, i.e., weights on choice options, that do not depend on the

Reproduced from: C. F. Schleussner et al., Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure, Sci. Rep., vol. 6, no. 1, p. 30790, Nov. 2016, doi: 10.1038/srep30790. Published under Creative Commons Attribution License 4.0 (CC BY).

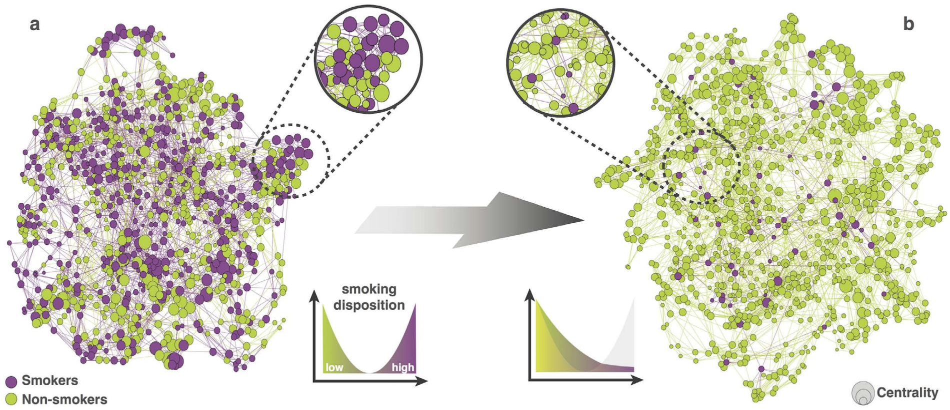


Figure 2. Smoking behaviour and centrality before and after the social transition. Panel (a,b) illustrate the initial and the final state of the contact network as simulated by the proposed adaptive network model of behaviour selection. Circles represent individual nodes. Their colour and size represent smoking behaviour and the individual's centrality, respectively. At the initial state of the simulation (panel (a)), smoking behaviour is homogeneously distributed across the network with random centrality values and the number of smokers equals the number of non-smokers resulting from the initial distribution of smoking dispositions. As the normative support for smoking gradually declined, behaviour and centrality changed over repeated interactions within the network. In the final state of the simulation (panel (b)), the number of smokers has considerably declined. As a consequence of the adaptive network dynamics, the centrality of smokers is selectively reduced. In comparison, non-smokers are characterised by a wide distribution of centrality.

interaction network. Such dispositions may be understood as culturally transmitted norms, values, knowledge and slowly changing collective contexts (e.g. health campaigns or climate change). Intuitively, by altering these weights according to a simulation protocol, one can emulate changes in global societally relevant factors. The hierarchical coupling between components in our model supports decomposition into partial models (see Table 1 and Methods). This allows us to differentiate the relative importance of model components and their associated social processes for behaviour selection in response to changing global trends.

Modelling Social Dynamics of Changing Smoking Behaviour

In the following, we apply the proposed adaptive network model of behaviour selection to the specific case of network-dependent changes in smoking behaviour to investigate empirically observed social transitions.

In particular, we introduce an update mechanism for the agent's smoking behaviour as the individual characteristic of interest. We conceptualise *smoking behaviour* as a binary variable (either smoking or non-smoking) endogenously in the model. The individual's smoking behaviour can be altered over time by an Ising-type model of social influence[13]. At each time step, we determine the probability of an agent to alter its smoking behaviour as a function of balanced peer-influence of smoking and non-smoking behaviour of its neighbours in the interaction network (see Methods). In addition to the peer-influence, we introduce an *individual smoking disposition* that reflects individual preferences in the probability to switch smoking behaviour. As this individual smoking disposition is exogenous to the model, its distribution can be altered externally and the dynamic response of the model can be investigated.

Importantly, it is only the endogenous binary smoking behaviour that dynamically affects agents' social proximity in our model. As in actual social networks, however, the social proximity also reflects many dimensions of which most remain latent during an interaction. Our *proximity matrix* thus includes two components: the time-invariant *background proximity* that largely determines the position of agents in the social network, e.g. reflecting long-term social ties such as family relationships, and a time-dependent component that depends on the co-occurrence of smoking behaviour for pairs of individuals. As a consequence, adopting a new behaviour will modify an agent's entries in the proximity matrix and may thus lead to changes in the contact network.

We then emulated dynamics of societal changes that historically lead to reduced prevalence of smoking (e.g. health campaigns and changes in public opinion[17]) by gradually modifying the exogenous distribution of smoking disposition. Over 1000 model time steps, we gradually converted a bimodal distribution, representing a balanced share of smokers and non-smokers in the network, into a quasi unimodal distribution favouring non-smoking attitudes as depicted in Fig. 2. We subsequently performed simulations over an ensemble of 1000 model runs using different seeds to initialise the pseudo-randomisation of the time-invariant background proximity matrix and the smoking disposition (see Methods). The model dynamics of interest were robust with respect to the specific choice of the distributions and the speed of the change in external forcing.

Results

To evaluate our co-evolutionary model of behaviour selection, we gradually modified the exogenous smoking disposition and studied the response of several metrics of our social adaptive network. These metrics were motivated

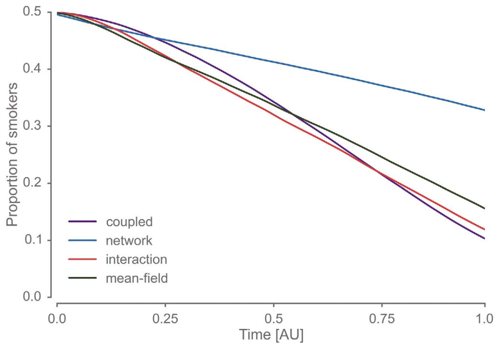


Figure 3. Gradual transition from a smoker to a non-smoker society is reflected in the prevalence of smoking. We considered four distinct models of behaviour selection. Over the course of the simulation, the distribution of the smoking disposition was gradually transformed from a bimodal to a quasi unimodal distribution and the smoking behaviour was computed at each time step. The *coupled* model assumes that behaviour is shaped by a local interaction based on a time-varying contact network. In the *network* model, no local interactions are considered. Here, the behaviour is only determined by the individual disposition while the contact network changes over time. In the *interaction* model, only local interactions shape the behaviour on a static contact network. Similarly to the *coupled* model, social influence and network dynamics are considered in the *mean-field* model, but smoking behaviour is shaped by non-local influences only. In all four (partial) models, the proportion of smokers changes in the course of the normative transition (time is represented in arbitrary units, AU). Notably, the absolute change is higher in the models that assume social interaction. Solid lines show mean values across 1000 runs with different pseudo-random initialisations. The variation across runs was negligible.

by previous empirical studies that documented co-evolution between behaviour and social network structure (cf. refs 17 and 19) and include the prevalence of smokers in the network, the eigenvector centrality of each individual and the probability that an individual smokes given that her contacts smoke (conditional probability of smoking).

Over time, the fraction of smokers in the network reduces from about 50% to 10% (Fig. 3), which is consistent with the empirical findings reported in ref. 17. In a second step, we compared different generative mechanisms by repeating the analysis for the remaining three partial models (see Table 1 and *coupled* model in Fig. 3). We found that models considering social influence dynamics (*interaction*, *mean-field* and *coupled*) reduced the smoking prevalence twice as much as the network model (Fig. 3), in which agent's behaviour is determined solely by the exogenous smoking disposition.

Only models considering the evolution of the contact network reduced the eigenvector centrality of remaining smokers below baseline (Fig. 4a). These effects were most pronounced for models combining social influence and network dynamics (*mean-field* and *coupled*). These models suggest a preferential reduction of centrality for smokers, reminiscent of the empirical results reported in ref. 17 (Fig. 4b). Here the *mean-field* model exhibits somewhat more drastic effects with less temporal variability as compared to the *coupled* model and even initially increases the eigenvector centrality of nodes, however not selectively for smokers. This is consistent with the deactivation of local influence that might give rise to "clusters of resistance". When considering the conditional probability of smoking up to fifth degree contacts (Fig. 5), we found changes between four to eight times higher in the *coupled* model as compared to all other models. This suggests that the feedback between specific *local* dynamics of social influence greatly amplifies such social clustering behaviour. Taken together, the results from our fully *coupled* co-evolutionary model support key findings from the empirical reference study[17]. At the same time, these results suggest that the residual pattern of clustered smokers of reduced centrality reflect a synergy between local interaction and network dynamics.

Discussion

We proposed a co-evolutionary model of behaviour selection in adaptive social networks and evaluated it through computational models targeting historical changes of smoking behaviour in social networks. Our computational models emulated gradual changes of network-wide smoking norms. We observed a reduced prevalence of smoking, a decreased eigenvector centrality and an increased conditional probability of smoking. Notably, the patterns of smoking behaviour and network characteristics computed by our model closely resemble empirical findings from a large-scale and long-term social network study investigating smoking behaviour in a North American small town[17]. Results of a partial model analysis suggest that selective modelling of either network dynamics, social influence or non-local social induction yields less match with empirical findings, and underscore the empirical relevance of behaviour-network co-evolution. Only the fully coupled co-evolutionary model was capable of explaining non-trivial structural change in complex social systems.

In particular, we would like to highlight the relevance of local generative mechanisms for social interaction as indicated by the deviating results for a *mean-field* forcing. The apparent imminent relevance of locality in interactions underscores the need for meaningful, social network based update mechanisms to study complex social phenomena.

Reproduced from: C. F. Schleussner et al., Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure, Sci. Rep., vol. 6, no. 1, p. 30790, Nov. 2016, doi: 10.1038/srep30790. Published under Creative Commons Attribution License 4.0 (CC BY).

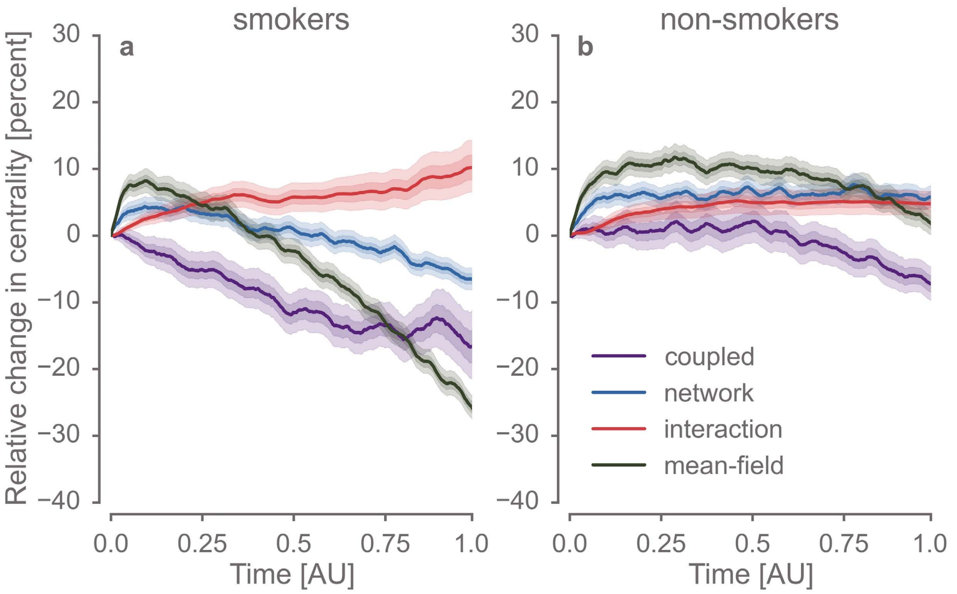


Figure 4. During the normative transition, local interactions between individuals and the evolution of the network structure rendered smokers less influential in the network. The eigenvector centrality (EVC) was computed at each time step for 1000 model runs with different random seeds. All models were initialised to the equilibrium run of the coupled model and values were normalised to the initial state of the model parameters. Panel (a,b) depict changes in EVC according to the four (partial) models for smokers and non-smokers, respectively. Solid lines show mean values and areas indicate bootstrapped 95% and 99% confidence intervals. It is noteworthy that only in models reflecting network effects, EVC was substantially reduced. These effects were strongest in models that in addition considered social interactions between individuals.

It is important to highlight that our models did neither involve any data-fitting nor predictive analyses. Instead we provided simulations with outputs according to their parameters and components. Thus the reported evidence is qualitative in nature and emphasises one distinct generative model through comparisons to empirical data and prior knowledge. The variance across ensembles assumes different values metric-wise, reflecting their algebraic properties as well as the effect of network size. Hence, our analyses do not imply statistical inference. The specific parameter choices in our model were adapted from the empirical study[17] and motivated from social sciences, evolutionary biology and neurosciences findings[2,19,52,59,60].

Furthermore, the models we evaluated in our simulations clearly suffer from conceptual limitations. The cognitive make-up of our simulated individual constitutes a bold simplification, particularly the assumption of an exogenously prescribed behavioural disposition. Human behaviour is clearly not binary but continuous and more complex model assumptions can therefore be easily motivated. Decision making is governed by multiple interacting factors, involving individual cognitive-emotional dispositions, but also by collaboration dynamics integrating social and cultural factors. Social support for a certain behaviour is often ambiguous, reflecting conflicting values, and social interactions can be asymmetric and unequally weighted. In this context, our model of behavioural change should be regarded as a prototype. We do not assess the validity of a specific mechanism of behavioural change or opinion formation. Instead we emphasise the structural importance of co-evolutionary processes that coalesce social cognition with network dynamics. But we hope that our simulation method stimulates future validation of specific social cognition theories against the background of evolving social networks. Nevertheless, our model generalises to other empirically documented examples of behaviour-network co-evolution including the spread of happiness, the spread of obesity but also the conditioning of food choices[61], as well as large data sets available from monitored social dynamics in massive multiplayer online games[20]. Assessing behavioural changes in social networks thereby complements spatial analysis[62], as social ties and physical distance tend to be substantially correlated[17] albeit a spatial and a network approach highlight fundamentally different qualities of the environment.

In particular, the example of food choices illustrates the potential outreach of our model for diverse interdisciplinary research questions. For example, the environmental foot-print of meat-centred diets is considerably higher than that of a vegetarian diet[63]. Against the historical background of strong positive correlations between meat consumption and economic prosperity[64] and given the rise of the global middle-class, diet habits represent a key challenge affecting several planetary boundaries[65,66]. At the same time, modifying nutritional behaviour has been targeted by health-related disciplines such as clinical psychology and behavioural medicine in preventive and therapeutic contexts. Capitalising on contingencies between individual behaviour and environment, therapeutic efforts might therefore benefit from models that specifically detail the relationship between microscopic

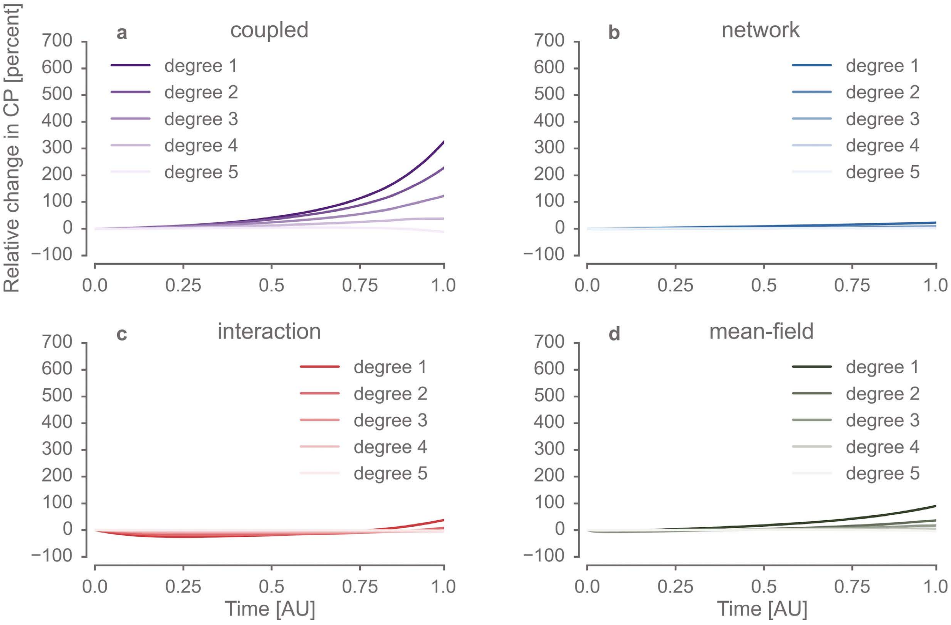


Figure 5. During the normative transition, joint effects of individual interaction and the evolution of the network structure led remaining smokers to cluster in marginalized groups. The probability of an individual being a smoker conditional on the prevalence of smoking in its neighbourhood at a given social distance or *degree* of separation (conditional probability, CP) was computed for 1000 model runs with different pseudo-random seeds at each time step and normalised to the percentage of change relative to the initial value. Panels (a–d) show the CP at social distance 1–5 for the *network*, the *interaction*, the *coupled* and the *mean field* models, respectively. CP increased with simulated time for lower social distances (1–3) across all models, whereas decreases of CP were observed for larger social distances in some models. Notably, this pattern was between four to eight times more pronounced in the *coupled* model compared to all other models. Solid lines depict mean values and areas represent bootstrapped 95% and 99% confidence intervals, barely visible as a result of the high signal to noise ratio for this metric.

and macroscopic social dynamics. Furthermore, co-evolutionary dynamics of behaviour and networks have been found to promote cooperation in public good games[67,68], thereby illustrating their transformative potential[69]. In the neuroscientific context it would be worthwhile to explore adaptive networks of spontaneous brain activity. Such models might help to overcome the limitations of ubiquitous “flat models” which do not resolve functional and structural connectivity hierarchically.

At a theoretical level, our study promotes a synergistic, co-evolutionary and interdisciplinary approach to social dynamics which gains explanatory momentum by integrating interpersonal cognitive processes with network dynamics as explanatory factors.

Methods

Detailed model description.

In the following, we provide a detailed mathematical model description, including definitions of the model variables and parameters, their initialisation, the algorithm for computing their temporal dynamics, the general modelling protocol and partial models. The model code is publically available and can be assessed and reviewed here: https://github.com/pik-copan/pycopanbehave. The implementation is based on the Python complex network software package pyunicorn[70] that is available at https://github.com/pik-copan/pyunicorn.

Model entities. We model individuals as *agents* or nodes $i \in V$, where V denotes the population or set of $N = |V|$ agents in the social system considered. The system is assumed to be closed and, hence, V does not depend on time, i.e. agents cannot enter or leave the population. The following entities define the system on the individual and population levels:

Agent properties. Each agent *i* carries a vector of scalar *agent properties*. Agent properties are parameters prescribed externally, they are fixed at initialisation and can be changed over time only by forcing external to the model (see below). The current model setup implements two agent properties:

(i) *Degree preference* $q_i \leq N-1$ is a discrete quantity serving as an upper bound of an agent’s degree in the contact network k_i^C (i.e. its number of neighbours in the contact network). This reflects the varying and limited capa-

Reproduced from: C. F. Schleussner et al., Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure, Sci. Rep., vol. 6, no. 1, p. 30790, Nov. 2016, doi: 10.1038/srep30790. Published under Creative Commons Attribution License 4.0 (CC BY).

bility of individuals to establish, manage and maintain sustained social relationships[2]. q_i is drawn from a discretised Gaussian distribution with mean μ and standard deviation σ at initialisation of the model. Specifically, we choose $\mu = 10$ and $\sigma = 3$ in the smoking experiment.

(ii) *Smoking disposition* $\gamma_i(t) \in [0, 1]$ is a continuous variable measuring an agent's individual and network-independent preference for smoking. Agents with small smoking disposition $\gamma_i(t)$ have a low probability to start smoking if they are non-smokers, while those with large $\gamma_i(t)$ have a low probability to stop smoking if they are smokers. $\gamma_i(0)$ is drawn at initialisation from a bimodal, parabolic probability density distribution that is optionally modified by stochastic external forcing towards a quasi unimodal distribution over time (see below).

Agent characteristics. Each agent i also carries a vector of scalar *agent characteristics*. The latter are dynamic variables that are by definition subject to social influence (induction), i.e. they are internal variables of the model obeying the social influence loop (Fig. 1). In this work, the following single agent characteristic is implemented: *smoking behaviour* $s_i(t) \in \{0, 1\}$ is a binary variable describing an agent's actual smoking behaviour. $s_i(t) = 0$ means that an agent does not smoke at time t, $s_i(t) = 1$ implies that an agent does smoke at time t.

Contact network. The *contact network* resembles the social relationships between agents. By contacts we refer explicitly to people that interact on a regular basis and are able to follow each other's affairs. We describe it as an undirected and simple time-dependent graph $G^C(t)$. It can be represented by its adjacency matrix $\mathbf{A}^C(t)$. The neighbourhood of agent i in the contact network is denoted by $\mathcal{N}_i^C(t)$.

Interaction network and interaction probability matrix. The *interaction network* represents all short-term interactions established between agents at each time step t. It is the basis for updating both the agent characteristics and the contact network at each time step. We describe it as an undirected and simple time-dependent graph $G^I(t)$. It can be represented by its adjacency matrix $\mathbf{A}^I(t)$. The neighbourhood of agent i in the interaction network is denoted by $\mathcal{N}_i^I(t)$.

Each entry $\pi_{ij}(t)$ of the *interaction probability matrix* $\Pi(t)$ gives the probability that two agents i, j will interact in time step t. For computing $\pi_{ij}(t)$, the social distance between two agents i, j is measured by the shortest path length $d_{ij}^C(t-1)$ between them in the contact network at time $t-1$ (the minimum number of steps needed to reach agent i from agent j over the contact network). Empirical results reveal that social influence decays approximately exponentially with d_{ij}^C[19]. Following these findings reporting a so-called "three degrees of separation law" of social influence, we define the interaction probability matrix as

$$\pi_{ij}(t) = (\beta - \varepsilon)\exp\left(-\frac{d_{ij}^C(t-1) - 1}{\delta}\right)\hat{L}(d_{ij}) + \varepsilon, \tag{1}$$

where ε is a baseline probability of interaction irrespective of the contact network. To account for the distribution of shortest path lengths between nodes, a normalisation factor $\hat{L}(d_{ij}) = L(1)/L(d_{ij})$ is introduced with $L(1)$ and $L(d_{ij})$ being the absolute number of shortest paths between nodes of length equal to 1 or d_{ij}, respectively. $\pi_{ij}(t)$ is scaled such that the probability of interaction between direct neighbours is always equal to the *interaction probability scaling factor* β. Here we set $\beta = 0.8$ and $\varepsilon = 0.03$. The parameter δ gives the typical social distance for the exponential decay of interaction probability and is chosen as $\delta = 2$ in line with empirical evidence[19].

Proximity matrix. The *proximity matrix* $\mathbf{P}(t)$ with elements $P_{ij}(t)$ measures the social proximity of two agents i, j. If social proximity is large, both agents are more likely to establish or maintain a contact. Hence, the proximity matrix is used in updating the contact network (see below). In our case of a single binary individual characteristic, the social proximity $P_{ij}(t)$ of two agents is only determined by the smoking behaviour $s_i(t)$ and $s_j(t)$ and their initially prescribed background proximity B_{ij} describing rigid social ties such as family relationships and other factors that are not explicitly included in the model. We choose a simple linear relationship

$$P_{ij}(t) = \alpha(1 - |s_i(t) - s_j(t)|) + (1 - \alpha)B_{ij}, \tag{2}$$

where α is a weight parameter balancing the influence of smoking behaviour and background proximity. Here, we choose $\alpha = 0.2$ to allow for a typical network mobility of between one and two degrees in the Watts-Strogatz network that underlies the background proximity generation.

The *background proximity matrix* $\mathbf{B}$ with elements B_{ij} is constructed on the basis of a Watts-Strogatz small-world network[45] with N nodes, mean degree $z = 10$ and a rewiring probability $p_w = 0.03$. The individual proximities B_{ij} are derived as a linear combination of the social distance d_{ij}^{WS} in a realisation of a Watt-Strogatz random network and a uniformly distributed stochastic component $\zeta \in [0, 1)$. This choice allows to emulate the typical small-world property of empirical social networks. B_{ij} is derived as

$$B_{ij} = 1 - 0.1(d_{ij}^{WS} - 1) - 0.1\zeta. \tag{3}$$

After computation of B_{ij} using the above formula, all entries with $B_{ij} < 0.2$ are reset to a minimum value of 0.2, which is in line with the assumption that for very high degrees of separation, no further meaningful distinction can be motivated.

Model dynamics. The temporal update scheme describes the dynamics of the main variables of interest in the model (Fig. 1): smoking behaviour $s_i(t)$ and contact network $G^C(t)$. The model evolves in discrete time steps. It is deterministic with the exception of the stochastic generation of the interaction network from the interaction probability matrix and the stochastic switching of the agents' smoking behaviours in each time step. After initialisation, the algorithm proceeds from step 1 to step 6 in the full co-evolutionary setup and then starts again at step 1. Modified model dynamics implemented to isolate the effects of specific mechanisms in the model are described below.

Step 1: Calculate interaction probabilities based on social distance. The interaction probability matrix Pi(t) is computed based on the social distance $d_{ij}^C(t-1)$ between agents i, j in the contact network $G^C(t-1)$ from the previous time step $t-1$ following Eq. 1.

Step 2: Generate interaction network. The interaction network's adjacency matrix is randomly generated for all $i, j \in V$ independently. An interaction takes place with probability $\pi_{ij}(t)$ corresponding to setting $A_{ij}^I(t) = 1$, while no interaction takes place with probability $1-\pi_{ij}(t)$ leading to $A_{ij}^I(t) = 0$.

Step 3: Change agent characteristics (social influence/induction step). In the considered case of a single binary individual characteristic (smoking behaviour), social influence reduces to a probabilistic switching of smoking behaviour similar to the flipping of spins in an Ising model in physics[13]. At time step t, the probability $p_i(t)$ to switch the smoking behaviour is assumed to depend to both the smoking disposition $\gamma_i(t-1)$ of agent i and the average smoking behaviour in the agent's neighbourhood in the interaction network $G^I(t)$ at time t. For all agents i, the smoking behaviour is determined as follows. For a non-zero number of interactions $k_i^I(t) = \sum_{j=1}^N A_{ij}^I(t)$:

- If $s_i(t-1)=0$ (non-smoker):

$$p_i(t) = C\gamma_i(t-1)\sum_{j\in\mathcal{N}_i^I(t)}\frac{s_j(t-1)}{k_i^I(t)} \tag{4}$$

The smoking behaviour switches to "smoker" with probability $p_i(t)$, i.e. $s_i(t)=1$, and remains the same with probability $1-p_i(t)$, i.e. $s_i(t)=0$.

- If $s_i(t-1)=1$ (smoker):

$$p_i(t) = C(1-\gamma_i(t-1))\sum_{j\in\mathcal{N}_i^I(t)}\frac{1-s_j(t-1)}{k_i^I(t)} \tag{5}$$

The smoking behaviour switches to "non-smoker" with probability $p_i(t)$, i.e. $s_i(t)=0$ and remains the same with probability $1-p_i(t)$, i.e. $s_i(t)=1$.
If no interactions take place for agent i ($k_i^I(t) = 0$), we set $s_i(t)=s_i(t-1)$.

The *smoking behaviour switching probability scaling factor C* scales the switching probability $p_i(t)$ of the smoking behaviour. C controls the amplitude of equilibrium stochastic noise of the smoking behaviour that is introduced by the Ising-like implementation. Here we set $C=0.1$.

Step 4: Calculate proximity matrix. The proximity matrix $\mathbf{P}(t)$ is computed from the current agent characteristics (smoking behaviour $s_i(t)$) and background proximity matrix $\mathbf{B}$ according to the diagnostic relationship given in Eq. 2.

Step 5: Update contact network. New ties in the contact network $G^C(t)$ can only be established between agents that interacted in the same time step. In contrast, potentially any edge may disappear from the contact network depending on the outcome of the following update scheme. Let $U_i(t)$ be the ordered set of neighbours $\mathcal{N}_i^I(t) \cup \mathcal{N}_i^C(t-1)$ of i included in its interaction and contact neighbourhoods that is sorted in descending order of social proximity $P_{ij}(t)$. The number of potential contacts of agent i is determined by the agent's degree preference q_i. Specifically, the agent's potential contact neighbourhood at time t consists of the set $T_i(t)$ of the first q_i entries of $U_i(t)$. Additionally, we require bidirectionality of contacts. This implies that only those agents can establish or maintain a contact relationship at time t that are included in each other's potential contact lists $T_i(t)$. Thus, the contact network at time t is derived as follows:

$$A_{ij}^C(t) = \begin{cases}1 & \text{if } i\in T_j(t)\wedge j\in T_i(t),\\ 0 & \text{otherwise.}\end{cases} \tag{6}$$

This means that a new contact relation can only be formed if the corresponding social proximity value is large enough to enter the potential contact neighbourhood of both involved agents. On the contrary, a contact is lost if either alter is not element of ego's own list of potential contacts or if ego is not element of alter's own list, or if both is the case. Thus ego can loose a contact actively (by dropping alter) or passively (by being dropped by alter).

Importantly, we do not derive "second best" solutions by iteratively updating the potential contacts after the bidirectionality check. We account for this refinement implicitly via the iterative network update dynamics cycle in the model (Fig. 1).

Step 6: Apply external forcing (optional). External forcing can change agent properties and other system parameters. In our smoking case study, we change the smoking dispositions $\gamma_i(t)$ of agents over time according to prescribed initial and target distributions to emulate the effects of changing values, political and health campaigns, etc. For all time steps, we ensure that the set $\{\gamma_i(t)\}_i$ is consistent with being drawn from a parabolic probability distribution of the form $y(x; t) = a(t)(b(t) - x)^2 + c(t)$ for $x \in [0, 1]$ with parameters implicitly defined by the conditions $\int_0^1 y(x;\, t)\, dx = 1$, $y(x = 0;\, t) = C_1$ and $y(x = 1;\, t) = C_2(t)$. Specifically, the initial $\{\gamma_i(0)\}_i$ are drawn from a bimodal and symmetric distribution $y(x; t = 0)$ with $C_2(t) = C_1$.

Then, the target distribution is changed over time by gradually reducing the parameter $C_2(t)$. To compute the set of smoking distributions $\{\gamma_i(t)\}_i$ at time step t, the previous set $\{\gamma_i(t - 1)\}_i$ is stochastically transformed by stepwise addition of two-tailed log-normal distributed noise ε that is linearly weighted by the deviation from the target distribution. Noise is added iteratively until a Kolmogorov-Smirnoff criterion with significance level 90% of $\{\gamma_i(t)\}_i$ being drawn from the target distribution $y(x; t)$ is fulfilled. By this procedure, individual γ_i are modified following a Markov process, whereas the overall system property, in this case the smoking disposition distribution, is externally controlled. Using this procedure, the randomly sampled initial set of smoking dispositions $\{\gamma_i(0)\}_i$ following a bimodal distribution ($C_2(t = 0) = C_1$) is gradually transformed into a sample $\{\gamma_i(t_f)\}_i$ following a quasi unimodal distribution ($C_2(t = t_f) = C_f$; see Fig. 1).

Modelling protocol. Model runs proceed in three steps: (i) The interaction network is initialised as $A_{ij}^{I}(0) = 1$ for all pairs i, j. In the following, the initial contact network $A_{ij}^{C}(0)$ is established based on the fully connected interaction network. Smoking behaviour $s_i(0)$ is initialised consistently with the initial smoking disposition $\gamma_i(0)$ as

$$s_i(0) = \Theta\left(\gamma_i(0) - \frac{1}{2}\right), \tag{7}$$

where $\Theta(\cdot)$ is the Heaviside function. (ii) The system is then integrated without applying external forcing for 200 time steps to a quasi-equilibrium state. We choose system parameters interaction probability scaling factor $\beta = 0.8$ and smoking behaviour switching probability scaling factor $C = 0.1$ to limit the system's internal noise level in equilibrium. More specifically, this choice guarantees that the maximum deviation from the median number of smokers in equilibrium that is induced by the stochastic dynamics of the model is smaller than 5% of the population size N. (iii) The system is then further integrated under continuous application of the stochastic external policy forcing acting on the smoking dispositions $\gamma_i(t)$.

Partial models. Besides the fully *coupled* model described above, we study three additional partial models that focus on a subset of processes of behaviour formation (Table 1): (i) an *interaction* model focussing on local social influence by assuming a static contact network (omitting step 5), (ii) a *network* model not considering local social influence, but inducing behavioural change only whenever the exogenously modified smoking disposition $\gamma_i(t)$ of an individual i crosses a threshold of 0.5 (modifying step 3), (iii) a *mean-field* model, where the agents react to the mean-field effect of the average smoking prevalence $S(t)/N$ instead of their local neighbourhood in the social influence process (modifying step 3). $S(t)$ is the number of smokers in time step t.

Network metrics. *Eigenvector centrality.* The eigenvector centrality $c_i(t)$ of agent i (also referred to simply as *centrality* above) is a non-local centrality measure implicitly defined to be proportional to the sum of i's contact neighbours' eigenvector centralities[71]. An agent has a high centrality if it is connected to many high centrality neighbours in the contact network that also have many high centrality neighbours. This implies that large values of eigenvector centrality $c_i(t)$ are observed in high density cliques or substructures embedded within the contact network. $c_i(t)$ is given by the i-th component of the leading eigenvector (associated to the largest eigenvalue) of the contact network's adjacency matrix $\mathbf{A}^C(t)$ at time t and is computed by applying the evcent method from the igraph package[72].

Conditional probability of smoking. The conditional probability that a randomly drawn agent i (ego) smokes given that another agent j (alter) randomly drawn from a neighbourhood shell at social distance $d_{ij}^{C}(t) = d$ smokes is defined as

$$P(d;\, t) = \frac{\frac{1}{S(t)} \sum_{i \in \mathcal{S}(t)} \frac{1}{N_{i,d}(t)} \sum_{j \in \mathcal{N}_{i,d}(t)} s_j(t)}{S(t)/N}. \tag{8}$$

Here, $\mathcal{S}(t)$ denotes the set of smokers at a given time-step and $S(t) = |\mathcal{S}(t)|$. Similarly, $\mathcal{N}_{i,d}(t)$ is the set of agents at a social distance d (measured by distance on shortest paths) from agent i in the contact network and $N_{i,d}(t) = |\mathcal{N}_{i,d}(t)|$ is the total number of agents in this set. In our study, $P(d; t)$ is employed as a measure of the mean effect of social distance in the contact network on smoking behaviour[17].

References

1. Skinner, B. F. Selection by consequences. *Science* **213,** 501–504 (1981).
2. Dunbar, R. I. M. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences* **16,** 681 (1993).
3. Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B. & Tomasello, M. Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science* **317,** 1360–1366 (2007).
4. Tomasello, M., Carpenter, M., Call, J., Behne, T. & Moll, H. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences* **28,** 675–691 (2005).

5. Fehr, E. & Camerer, C. F. Social neuroeconomics: The neural circuitry of social preferences. *Trends in Cognitive Sciences* **11,** 419–427 (2007).
6. Engemann, D. A., Bzdok, D., Eickhoff, S. B., Vogeley, K. & Schilbach, L. Games people play–toward an enactive view of cooperation in social neuroscience. *Frontiers in Human Neuroscience* **6** (2012).
7. Festinger, L. *A theory of cognitive dissonance*, vol. 2 (Stanford university press, 1962).
8. Janis, I. L. & Mann, L. *Decision making: A psychological analysis of conflict, choice, and commitment*. (Free Press, 1977).
9. Schelling, T. C. Dynamic models of segregation†. *Journal of Mathematical Sociology* **1,** 143–186 (1971).
10. Axelrod, R. The dissemination of culture a model with local convergence and global polarization. *Journal of Conflict Resolution* **41,** 203–226 (1997).
11. Akerlof, G. A. Social distance and social decisions. *Econometrica* **65,** 1005–1027 (1997).
12. Steinbacher, M., Steinbacher, M. & Steinbacher, M. *Interaction-based approach to economics and finance*, 161–203 (Springer International Publishing, Cham, 2014).
13. Castellano, C., Fortunato, S. & Loreto, V. Statistical physics of social dynamics. *Reviews of Modern Physics* **81,** 591 (2009).
14. Newman, M. *Networks: an introduction* (Oxford University Press, 2010).
15. Christakis, N. A. & Fowler, J. H. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine* **357,** 370–9 (2007).
16. Fowler, J. H., Christakis, N. A. *et al.* Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *British Medial Journal* **337,** a2338 (2008).
17. Christakis, N. A. & Fowler, J. H. The collective dynamics of smoking in a large social network. *The New England Journal of Medicine* **358,** 2249–58 (2008).
18. McDermott, R., Fowler, J. H. & Christakis, N. A. Breaking up is hard to do, unless everyone else is doing it too: Social network effects on divorce in a longitudinal sample. *Social Forces* **92,** 491–519 (2013).
19. Christakis, N. A. & Fowler, J. H. Social contagion theory: Examining dynamic social networks and human behavior. *Statistics in Medicine* **32,** 556–77 (2012).
20. Szell, M. & Thurner, S. Measuring social dynamics in a massive multiplayer online game. *Social Networks* **32,** 313–329 (2010).
21. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 7821–6 (2002).
22. Asch, S. E. Opinions and social pressure. *Readings about the Social Animal* **193,** 17–26 (1955).
23. Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. & Suri, S. Feedback effects between similarity and social influence in online communities. *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD' 08* 160 (2008).
24. Colman, A. M. Cooperation, psychological game theory, and limitations of rationality in social interaction. *Behavioral and Brain Sciences* **26,** 139–153 (2003).
25. Snijders, T. A., Van de Bunt, G. G. & Steglich, C. E. Introduction to stochastic actor-based models for network dynamics. *Social Networks* **32,** 44–60 (2010).
26. Gross, T. & Blasius, B. Adaptive coevolutionary networks: a review. *Journal of The Royal Society Interface* **5,** 259–271 (2008).
27. Gross, T. & Sayama, H. (eds.) *Adaptive networks* (Springer, Berlin Heidelberg, 2009).
28. Sayama, H. *et al.* Modeling complex systems with adaptive networks. *Computers & Mathematics with Applications* **65,** 1645–1664 (2013).
29. Holme, P. & Newman, M. E. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E* **74,** 056108 (2006).
30. Diakonova, M., Eguluz, V. M. & San Miguel, M. Noise in coevolving networks. *Physical Review E* **92,** 032803 (2015).
31. Gross, T., D'Lima, C. J. D. & Blasius, B. Epidemic dynamics on an adaptive network. *Physical Review Letters* **96,** 208701 (2006).
32. Huepe, C., Zschaler, G., Do, A.-L. & Gross, T. Adaptive-network models of swarm dynamics. *New Journal of Physics* **13,** 073022 (2011).
33. Li, M. *et al.* A coevolving model based on preferential triadic closure for social media networks. *Scientific Reports* **3,** 2512 (2013).
34. Wiedermann, M., Donges, J. F., Heitzig, J., Lucht, W. & Kurths, J. Macroscopic description of complex adaptive networks coevolving with dynamic node states. *Physical Review E* **91,** 052801 (2015).
35. Auer, S., Heitzig, J., Kornek, U., Schöll & Kurths, J. The dynamics of coalition formation on complex networks. *Scientific Reports* **5,** 13386 (2015).
36. Nardini, C., Kozma, B. & Barrat, A. Who's talking first? Consensus or lack thereof in coevolving opinion formation models. *Physical Review Letters* **100,** 158701 (2008).
37. Demirel, G., Prizak, R., Reddy, P. N. & Gross, T. Cyclic dominance in adaptive networks. *European Physical Journal B* **84,** 541–548 (2011).
38. Böhme, G. A. & Gross, T. Analytical calculation of fragmentation transitions in adaptive networks. *Physical Review E* **83,** 035101 (2011).
39. Bozon, M. & Heran, F. Finding a spouse: A survey of how french couples meet. *Population English Selection No.* **1** 91–121 (1989).
40. Kossinets, G. & Watts, D. J. Empirical analysis of an evolving social network. *Science* **311,** 88–90 (2006).
41. Henry, A., Pralat, P. & Zhang, C. Emergence of segregation in evolving social networks. *Proceedings of the National Academy of Sciences of the United States of America* **108,** 8605 (2011).
42. Hegselmann, R. & Krause, U. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artifical Societies and Social Simulation (JASSS)* **5** (2002).
43. Van Essen, D. C., Anderson, C. H. & Felleman, D. J. Information processing in the primate visual system: An integrated systems perspective. *Science* **255,** 419–423 (1992).
44. Roberts, J. A., Boonstra, T. W. & Breakspear, M. The heavy tail of the human brain. *Current Opinion in Neurobiology* **31,** 164–172 (2015).
45. Watts, D. J. & Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature* **393,** 440–2 (1998).
46. Dehaene, S., Sergent, C. & Changeux, J.-P. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proceedings of the National Academy of Sciences of the United States of America* **100,** 8520–8525 (2003).
47. Flack, J. C. & Krakauer, D. C. Challenges for complexity measures: A perspective from social dynamics and collective social computation. *Chaos* **21,** 037108 (2011).
48. Flack, J. C. Multiple time-scales and the developmental dynamics of social systems. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **367,** 1802–1810 (2012).
49. DeDeo, S. Collective phenomena and non-finite state computation in a human social system. *PloS One* **8,** e75818 (2013).
50. Luhmann, C. C. & Rajaram, S. Memory transmission in small groups and large networks an agent-based model. *Psychological Science* **26,** 1909–1917 (2015).
51. DeDeo, S., Krakauer, D. C. & Flack, J. C. Inductive game theory and the dynamics of animal conflict. *PLoS Computational Biology* **6,** e1000782 (2010).
52. Domingue, B. W., Fletcher, J., Conley, D. & Boardman, J. D. Genetic and educational assortative mating among us adults. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 7996–8000 (2014).

Reproduced from: C. F. Schleussner et al., Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure, Sci. Rep., vol. 6, no. 1, p. 30790, Nov. 2016, doi: 10.1038/srep30790. Published under Creative Commons Attribution License 4.0 (CC BY).

53. Werner, C. & Parmelee, P. Similarity of activity preferences among friends: Those who play together stay together. *Social Psychology Quarterly* 62–66 (1979).
54. Eiser, J. R., Morgan, M., Gammage, P., Brooks, N. & Kirby, R. Adolescent health behaviour and similarity-attraction: Friends share smoking habits (really), but much else besides. *British Journal of Social Psychology* **30,** 339–348 (1991).
55. Kobus, K. Peers and adolescent smoking. *Addiction* **98,** 37–55 (2003).
56. Bickart, K. C., Hollenbeck, M. C., Barrett, L. F. & Dickerson, B. C. Intrinsic amygdala-cortical functional connectivity predicts social network size in humans. *The Journal of Neuroscience* **32,** 14729–14741 (2012).
57. Deary, I. J., Penke, L. & Johnson, W. The neuroscience of human intelligence differences. *Nature Reviews Neuroscience* **11,** 201–211 (2010).
58. Onnela, J.-P. & Reed-Tsochas, F. Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 18375–80 (2010).
59. Buzsáki, G. & Mizuseki, K. The log-dynamic brain: How skewed distributions affect network operations. *Nature Reviews Neuroscience* **15,** 264–278 (2014).
60. Linkenkaer-Hansen, K., Nikouline, V. V., Palva, J. M. & Ilmoniemi, R. J. Long-range temporal correlations and scaling behavior in human brain oscillations. *The Journal of Neuroscience* **21,** 1370–1377 (2001).
61. Pachucki, M. A., Jacques, P. F., Christakis, N. A., Wood, R. & Health, J. Social network concordance in food choice among spouses, friends, and siblings. *American Journal of Public Health* **101,** 2170–2177 (2011).
62. Gallos, L. K., Barttfeld, P., Havlin, S., Sigman, M. & Makse, H. A. Collective behavior in the spatial spreading of obesity. *Scientific Reports* **2** (2012).
63. Stehfest, E. *et al.* Climate benefits of changing diet. *Climatic Change* **95,** 83–102 (2009).
64. Popp, A., Lotze-Campen, H. & Bodirsky, B. Food consumption, diet shifts and associated non-CO2 greenhouse gases from agricultural production. *Global Environmental Change* **20,** 451–462 (2010). Governance, Complexity and Resilience.
65. Rockström, J. *et al.* A safe operating space for humanity. *Nature* **461,** 472–475 (2009).
66. Steffen, W. *et al.* Planetary boundaries: Guiding human development on a changing planet. *Science* **347,** 1259855 (2015).
67. Fowler, J. H. & Christakis, N. A. Cooperative behavior cascades in human social networks. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 5334–8 (2010).
68. Fehl, K., van der Post, D. J. & Semmann, D. Co-evolution of behaviour and social network structure promotes human cooperation. *Ecology Letters* **14,** 546–551 (2011).
69. Benn, S., Dunphy, D. & Griffiths, A. *Organizational change for corporate sustainability* (Routledge, 2014).
70. Donges, J. F. *et al.* Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package. *Chaos* **25,** 113101 (2015).
71. Bonacich, P. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology* **2,** 113–120 (1972).
72. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Systems CX.* **18,** 1695 (2006).

Acknowledgements

The reported research was conducted within the scope of the COPAN flagship project on co-evolutionary pathways at the Potsdam Institute for Climate Impact Research (http://www.pik-potsdam.de/copan). The authors acknowledge support by the German National Academic Foundation. D.A.E. is supported by the ERC StG 263584 awarded to Virginie van Wassenhove, the Amazon in Education Grant awarded to D.A.E., as well as the INSERM, CEA, and University Paris Sud. J.F.D. thanks the Stordalen Foundation (via the Planetary Boundary Research Network PB.net) and the Earth League's EarthDoc program for financial support. The authors gratefully acknowledge the European Regional Development Fund (ERDF), the German Federal Ministry of Education and Research and the Land Brandenburg for supporting this project by providing resources on the high performance computer system at the Potsdam Institute for Climate Impact Research. The publication of this article was funded by the Open Access Fund of the Leibniz Association. Danilo Bzdok, Guillaume Dumas, Jobst Heitzig, Reik V. Donner, Wolfram Barfuss and Virginie van Wassenhove are acknowledged for valuable comments and discussions.

Author Contributions

C.-F.S., J.F.D., D.A.E. and A.L. designed the research. C.-F.S., J.F.D. and D.A.E. performed the research and wrote the manuscript. All authors reviewed the manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Schleussner, C.-F. *et al.* Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure. *Sci. Rep.* **6**, 30790; doi: 10.1038/srep30790 (2016).

LETTERS
https://doi.org/10.1038/s41558-018-0079-z
nature climate change
Corrected: Publisher Correction

Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility

Jobst Heitzig[1]* and Ulrike Kornek[1,2]

The Paris Agreement relies on nationally determined contributions to reach its targets and asks countries to increase ambitions over time, leaving open the details of this process. Although overcoming countries' myopic 'free-riding' incentives requires cooperation, the global public good character of mitigation makes forming coalitions difficult. To cooperate, countries may link their carbon markets[1], but is this option beneficial[2]? Some countries might not participate, not agree to lower caps, or not comply to agreements. While non-compliance might be deterred[3], countries can hope that if they don't participate, others might still form a coalition. When considering only one coalition whose members can leave freely, the literature following the publication of refs [4,5] finds meagre prospects for effective collaboration[6]. Countries also face incentives to increase emissions when linking their markets without a cap agreement[7,8]. Here, we analyse the dynamics of market linkage using a game-theoretic model of far-sighted coalition formation. In contrast to non-dynamic models and dynamic models without far-sightedness[9,10], in our model an efficient global coalition always forms eventually if players are sufficiently far-sighted or caps are coordinated immediately when markets are linked.

Our study extends the climate coalition literature by analysing a dynamic process with multiple coalitions, far-sighted players anticipating further steps, and uncertainty about which transitions will happen[11–13] (in contrast to cost and benefit uncertainty). We adapt the dynamic far-sighted coalition formation model of ref. [13] to the linking of carbon markets with endogenous decisions whether to coordinate caps. Unlike refs[14–17] which focus on stable end results, our model allows insights about the process. We assume these dynamic possibilities:

1. Individual countries or regions establish carbon markets to cost-efficiently achieve individual mitigation goals.
2. Market linkage: some markets get linked to reduce costs by equalizing marginal abatement costs, leading to adjustments in members' emissions caps (for example, refs [18,19]).
3. Cap coordination: members of linked markets may agree to coordinate the amounts of permits each member issues, internalizing the effect of their emissions on each other, thus reducing their total cap[20,21]. This coordination may or may not already be part of the linkage agreement. Any agreement may be terminated at any time by any member.

This may eventually lead to a (near-)global emissions trading scheme with coordinated caps and substantial mitigation levels. Although first steps along this line have been taken already[22], it is unclear which markets will be linked, which caps coordinated, in which order, and whether this will lead to a global market with an efficient cap. We present scenarios of how the dynamic formation of linked carbon markets with coordinated caps might evolve.

In our model, a set of players can form and later terminate different markets and cap-coordinating coalitions over time (rectangular nodes in Fig. 1). Each constellation (for example, the constellation [AB],C, where players A and B are in an immediately coordinated market without player C) is a possible state x of the process and would result in certain static payoffs $\pi_i(x)$ if it would prevail (for example, Fig. 1c, middle column).

We use different settings for these static payoffs, at first a simple illustrative cost-benefit structure with linear benefits and marginal mitigation costs, then later a version of the coalitional payoffs from ref.[7] based on cost-benefit estimates from refs[23,24], assuming that surpluses from forming a coalition are shared according to the asymmetric Nash bargaining solution[25], that is, in proportion to some distribution of bargaining power, see 'Derivation of static payoffs' in the Methods section.

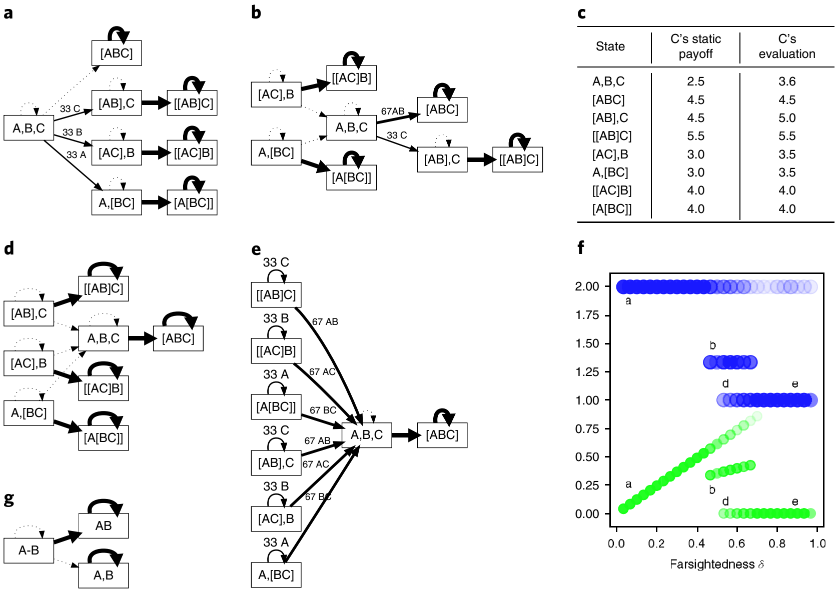
The possible transitions between states x,y,... (arrows in Fig. 1) represent the formation of new markets or coalitions (for example, adding an overarching three-player coalition to [AB], resulting in a transition from [AB],C to [[AB]C], Fig. 1b), or the termination of existing ones by some or all members (for example, the transition from [AB],C to the non-cooperative state A,B,C in Fig. 1e). Players hold beliefs about the process that are represented as subjective transition probabilities (shown as percentages) $p_{x\to y}$.

Given any assumed transition probabilities, a player i evaluates each state x by the discounted long-term payoffs $\ell_i(x)$ she can expect when starting in that state and then progressing according to these probabilities (Fig. 1c, right column). Our main parameter is the level of far-sightedness δ used in evaluations, representing the combined effects of time preference, trust that the process does not break down, and duration between steps. Mathematically,

$$\ell_i(x) = (1-\delta)\pi_i(x) + \delta \sum_y p_{x\to y}\ell_i(y). \tag{1}$$

At the same time, given any such evaluations, certain transitions appear unprofitable since they decrease the evaluation of some relevant player (for example, the transition A,B,C → [AC],B in Fig. 1b,c is unprofitable for player C in view of her beliefs about the further steps, although temporarily her payoffs would increase). A profitable transition is dominated if some of its relevant players can initiate another transition that they all prefer. In each step, a player is drawn at random with probabilities proportional to bargaining power. She proposes her favourite profitable and undominated transition (marked by arrow labels), and all relevant players

[1]Potsdam Institute for Climate Impact Research, Potsdam, Germany. [2]Mercator Research Institute on Global Commons and Climate Change, Berlin, Germany. *e-mail: heitzig@pik-potsdam.de



State	C's static payoff	C's evaluation
A,B,C	2.5	3.6
[ABC]	4.5	4.5
[AB],C	4.5	5.0
[[AB]C]	5.5	5.5
[AC],B	3.0	3.5
A,[BC]	3.0	3.5
[[AC]B]	4.0	4.0
[A[BC]]	4.0	4.0

Fig. 1 | Illustration of the model in a fictitious situation with three symmetric players A, B, and C, linear mitigation benefits, and quadratic mitigation costs. All eight possible coalition states are shown. **a**, Unique equilibrium process for low far-sightedness $\delta = 0.3$; arrow labels state transition probabilities in percent and which players favour this transition. **b**, One of three alternative equilibrium processes for medium far-sightedness $\delta = 0.5$. **c**, Static payoffs (in arbitrary units) and evaluations of player C in process **b**, based on linear benefits and marginal costs (see text). **d**, Unique result for high far-sightedness $\delta = 0.7$. **e**. Very high far-sightedness $\delta = 0.9$. **f**, Effect of far-sightedness on mean number of steps to reach a grand coalition (large dots) and on total payoff uncertainty (small dots, arbitrary units, see Methods), one dot for each existing equilibrium process, with dots' opacity indicating how often this process was found by our algorithm (see Supplementary Section 3.6 for details). **g**, Example where two asymmetric players can get stuck when immediate cap coordination is unavailable ($\delta = 0.5$, see Supplementary Section 3.2 for details).

accept this since they profit from it and cannot initiate a better transition. Note that she may propose a coalition that excludes herself (for example, because of fairness and responsibility). If an undominated profitable transition is no player's favourite, it gets zero probability (dotted arrows). This process of rationally proposing and accepting transitions generates a set of objective transition probabilities, which are thus a function of the given evaluations,

$$\{p_{x\to y}\} = f(\{\ell_i(x)\}), \tag{2}$$

and which can then be compared to the subjective probabilities the players started with.

If objective and subjective probabilities coincide, they describe an equilibrium process since they form a 'consistent' set of common beliefs that prove to be correct if all players act rationally with respect to these beliefs. In other words, an equilibrium is given if the two (typically large) systems of equations (1), (2) between all the quantities $p_{x\to y}$ and $\ell_i(x)$ are fulfilled. We identify such equilibrium processes numerically.

Consider the illustrative example of Fig. 1, where three symmetric players can form coalitions with static payoffs based on ref. [26]. Player i's benefits and costs from mitigating q_i units of greenhouse gas (GHG) emissions are $\sum_j q_j$ and $q_i^2/2$, respectively. For simplicity, let us assume for now that when forming a market the players must immediately agree on caps. For low to medium far-sightedness ($\delta < 0.45$) there is only one equilibrium process, where each player proposes that the other two form a coalition first before she joins (Fig. 1a). For $0.45 < \delta < 0.67$, there are three more alternative equilibrium processes in which all players believe that one of them (for example, C in Fig. 1b,c) would not join a bilateral coalition, resulting in a 2/3 probability of forming the

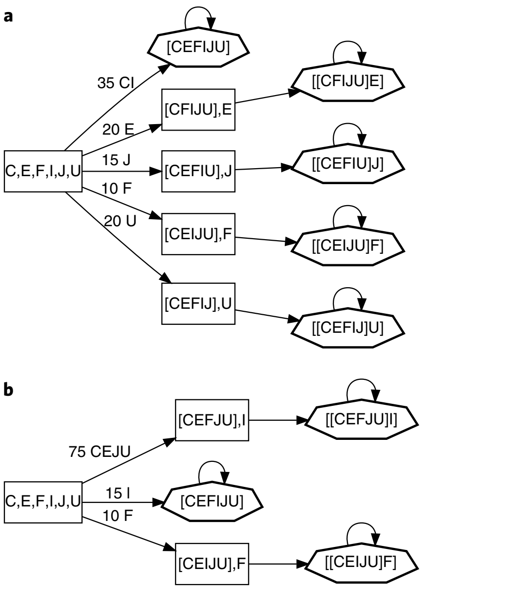


Fig. 2 | Typical model results for the six major emitters. a, Low to moderate far-sightedness (here $\delta = 0.5$). A fully coordinated global carbon market results after one step (with 35% probability, if the permit sellers C(hina) and I(ndia) get their way) or two steps (with 65% probability, if E(urope), F(ormer Soviet Union), J(apan), or U(SA) manage to stay out of the market at first). Diamond-shaped nodes are stable states with an optimal global cap, differing only in the burden- or surplus-sharing between members. See Table 1 for payoffs. **b**, Highly far-sighted players (here $\delta = 0.9$).

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility, J. Heitzig and U. Kornek, doi: 10.1038/s41558-018-0079-z, Copyright 2018, The Author(s).

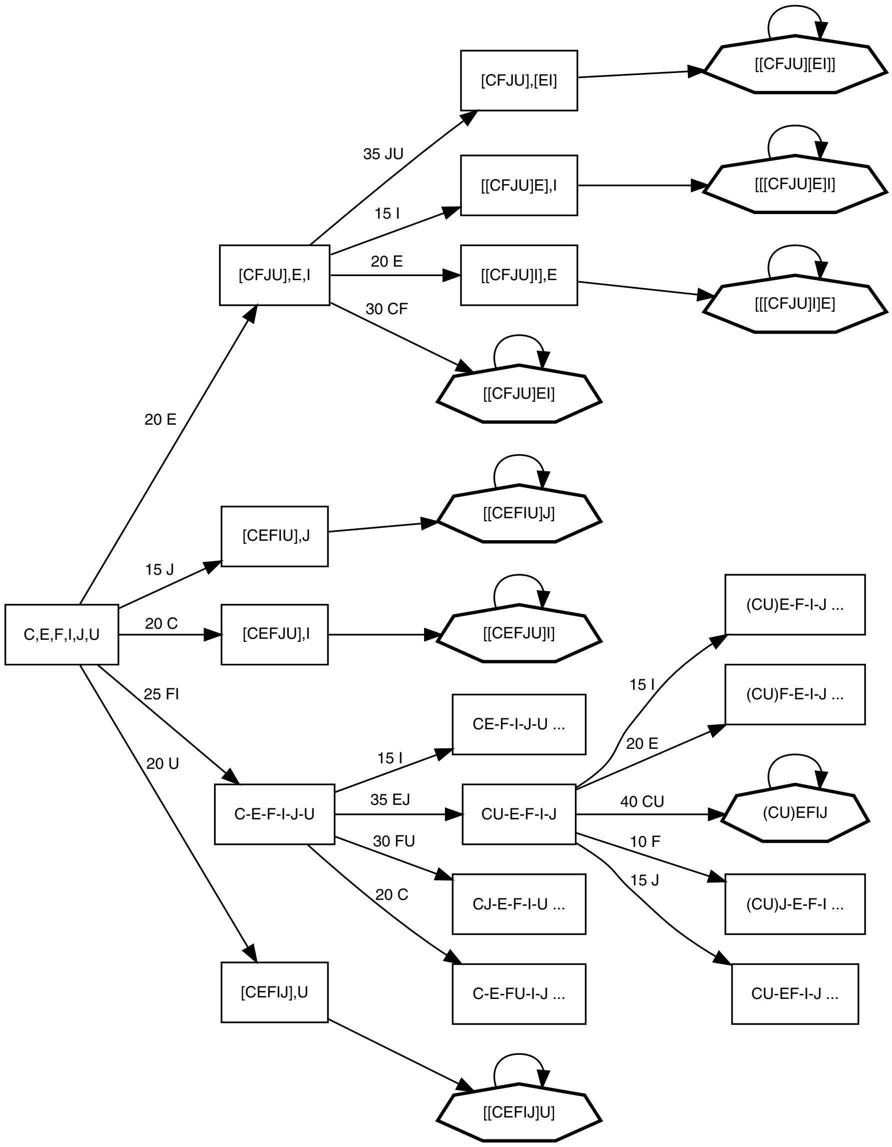


Fig. 3 | Alternative scenario to Fig. 2 with typical complications occurring if players are highly far-sighted ($\delta = 0.9$) and agreements are irreversible. In view of the expected later moves, F and I now prefer to establish a global market C-E-F-I-J-U that only later coordinates its caps and in which all members prefer to join cap coordination late. In that branch, only the path with the highest probability is shown completely here, ending in a fully coordinated market (CU)EFIJ in which C and U have formed a cap coordinating coalition first before agreeing with the others to coordinate further; other paths are pruned for the figure (marked by '…'). See Supplementary Table 1 for payoffs and evaluations and Supplementary Section 3.3 for a discussion.

grand coalition right away. For $0.53 < \delta < 0.75$, there is another equilibrium process (Fig. 1d) where no player can hope to stand back when starting with no collaboration in state A,B,C; in that equilibrium, however, players believe that if a bilateral coalition already exists for whatever reason (as in [AB],C), it would not be terminated but another overarching coalition would be formed (for example, [[AB]C]). Finally, for $\delta > 0.75$, this belief would become inconsistent with the evaluations since the two players in the bilateral coalition would become far-sighted enough to prefer terminating their coalition, anticipating the eventually higher payoffs in [ABC] (Fig. 1e). While for each given type of equilibrium, increasing far-sightedness increases the uncertainty about the resulting path, it overall reduces this uncertainty due to the change in which equilibria exist, and it makes it more likely that a grand coalition forms in just one step (Fig. 1f).

Also consider shortly the case where two players A,B cannot form a coordinated market [AB] in one step but need to first form an uncoordinated market, denoted A-B, and then agree on caps afterwards, denoted AB. Then, if δ is not large enough and vulnerabilities and cost efficiencies are asymmetric but considerably positively correlated, the first move may not be profitable and the unique equilibrium process may look like in Fig. 1g, remaining in the uncooperative state A,B. (see Supplementary Section 3.2 for an analysis).

For a more realistic picture, we identified equilibrium processes for a setting in which the static payoffs for the six major GHG

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility, J. Heitzig and U. Kornek, doi: 10.1038/s41558-018-0079-z, Copyright 2018, The Author(s).

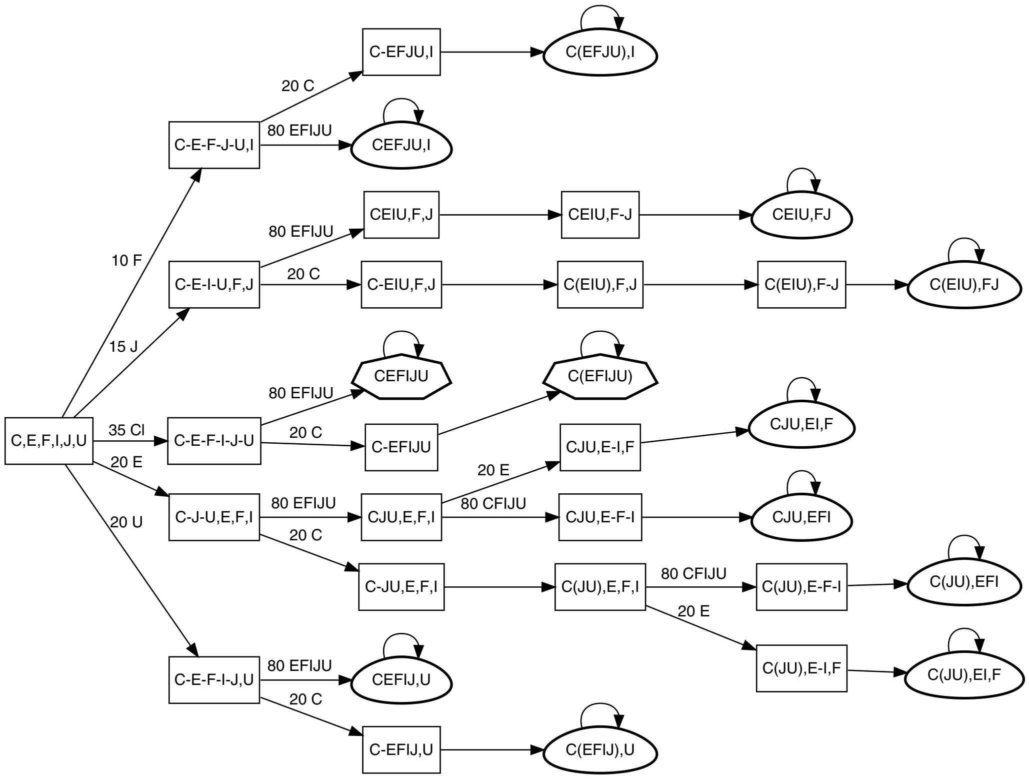


Fig. 4 | Alternative scenario to Fig. 2 in a world where caps cannot immediately be coordinated when markets are linked but only later in separate moves. Medium far-sightedness of $\delta = 0.5$, unilaterally terminable agreements. A fully coordinated global carbon market is only reached with 35% probability, otherwise the process gets stuck with two or more markets (egg-shaped nodes) since players are not far-sighted enough to accept the temporary costs of delayed cap coordination. (See Supplementary Figs. 3, 4 for irreversible agreements and myopic players.).

emitters C(hina), E(urope), F(ormer Soviet Union), I(ndia), J(apan), and U(SA) are derived from the literature (refs [7,23,24], see Methods), resulting in scenarios such as those depicted in Figs. 2,3 and 4. Table 1 compares the (myopic) static payoffs and (far-sighted) evaluations in some states of Fig. 2a. There, if the US were myopic, they would not consider forming an uncoordinated market with Japan resulting in a move from the initial state labelled (a) to (b) in Table 1, since if this state were to prevail, their payoff would be reduced (middle column U). Farsightedly, however, they anticipate further steps resulting in larger markets and an eventual increase in payoff (last column U), making the move (a) → (b) profitable after all.

Despite the strong dependency of actual transition probabilities on the model parameters, a systematic analysis of the above three-player case and the more realistic six-player setting reveals the following findings (see Supplementary Section 3 for details):

- If it is possible to immediately coordinate caps when linking markets, a global market with a first-best cap emerges, but probably not in one move (Fig. 2), and with uncertainty about who will cooperate first.
- Counterintuitively, when agreements are reversible (can be terminated), the process takes fewer steps and is less uncertain since agreements which would later be terminated are not signed in the first place (compare Figs. 2b and 3), so that no agreement actually signed will be terminated later. Higher far-sightedness tends to reduce uncertainty and the mean number of steps further (Figs. 1,2).
- When agreements are irreversible, a large market might be established at first with uncoordinated caps, which then eventually get fully coordinated in several further moves (for example, the "C-E-F-I-J-U" branch in Fig. 3). Higher far-sightedness here tends to increase uncertainty and the mean number of steps (Fig. 3) since it makes more and smaller transitions profitable.
- If immediate cap coordination is not an option when linking markets, and players are not sufficiently far-sighted, a global market may not emerge (as in Fig. 1g) and they might get stuck with several, only internally coordinated carbon markets (Fig. 4).

While these findings appear robust under simple variations such as further restricting the number of players and varying the cost, vulnerability, and bargaining power coefficients, the following effects may depend on the assumed linearity of benefits. First, free-riding by not entering a market: even when joining a market eventually, prospective permit buyers tend to have an incentive to free-ride by joining late, while prospective sellers tend to profit from joining early (for example, compare favourite moves and payoffs of C, seller, and U, buyer, in Fig. 2a and Table 1). A permit seller might or might not prefer if its main competitor joins the market only later (for example, compare the evaluations for C in state [CEFJU],I in Table 1 and Supplementary Section 2, and C's favourite moves in states C,E,F,I,J,U and [CFJU],E,I of Fig. 3). Second, free-riding by not coordinating caps: in a not yet fully coordinated market, both permit buyers and sellers usually have an incentive to free-ride by entering a coalition late (for example, in the C-E-F-I-J-U branch in Fig. 3). Overall, the analysis in Supplementary Section 3.6 shows

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility, J. Heitzig and U. Kornek, doi: 10.1038/s41558-018-0079-z, Copyright 2018, The Author(s).

Table 1 | Static payoffs derived from refs [7,23,24] and evaluations in the process shown in Fig. 2a for a typical choice of parameters

State	Static payoffs						Evaluations					
	C	E	F	I	J	U	C	E	F	I	J	U
C,E,F,I,J,U (a)	94	394	115	86	304	347	254	609	200	206	458	555
[CEFIJU]	484	785	310	379	597	738	*484	785	310	*379	597	738
[CFIJU],E	330	1182	233	263	481	584	352	*1204	244	280	498	606
[CEIJU],F	421	721	378	331	550	675	443	743	*389	348	566	696
[CEFIU],J	375	676	255	297	960	629	385	686	260	305	*968	639
[CEFIJ],U	326	626	231	260	478	1055	357	658	246	284	502	*1087
C-E-F-I-J-U	180	>360	231	217	321	>338	380	≫510	306	329	≫434	≫488
C-E-F-I-J,U	147	>381	189	169	317	439	≫237	≫466	232	230	≫378	659
C,E,F,I,J-U (b)	91	385	112	84	306	>338	175	480	160	147	512	607
C,[EFIJU]	215	565	200	214	433	519	329	680	258	300	519	633
[CFJU],E,I	298	1020	217	219	457	551	344	1078	240	255	492	597
[CEJU],F,I	381	681	328	245	520	635	430	730	354	282	556	684
[CEFJU],I	443	743	289	280	566	696	470	771	303	301	587	724
[CU],E,F,I,J	164	677	195	146	512	418	254	808	245	214	606	≫523
[[CFIJU]E]	374	1226	255	296	514	628	374	1226	255	296	514	628
[[CEIJU]F]	464	765	400	364	582	718	464	765	400	364	582	718
[[CEFIU]J]	395	696	265	312	975	649	395	696	265	312	975	649
[[CEFIJ]U]	389	689	262	307	526	1119	389	689	262	307	526	1119

Medium far-sightedness $\delta = 0.5$, unilaterally terminable agreements. Estimates are given in US$ billion per 100 years. Only states reached with positive probability (bold) and some alternative states are shown. *Favourite undominated move of this player in state C,E,F,I,J,U. > Indicates that move is not statically profitable for this initiating player. ≫ Indicates that move is not long-term profitable for this initiating player. (a),(b) are the states referred to in the main text.

that the combined effects of differences in vulnerability, cost efficiency, and bargaining power are highly nonlinear and can be very complicated.

Our scenarios show that an explicit modeling of the stepwise process of forming, merging and potentially terminating multiple coalitions of various size changes the often pessimistic picture of previous literature on coalition formation. Most importantly, while a country may have an incentive to delay cooperation to temporarily profit from others' efforts of cooperation, and thus improve their bargaining position for later steps, this 'free-riding' will not remove the incentive to later join an overarching coalition as long as further cooperation generates some surplus that the existing coalition can share with this country. In other words, the dynamic analysis shows that free-riding does not prevent the eventual formation of a grand coalition, but only changes the surplus (or burden) sharing within the grand coalition to the advantage of the free-riding country. Our model results thus give an alternative explanation of the currently observed low level of cooperation in international climate policy: rather than planning to free-ride permanently, some countries may at present try to stand back simply to improve their bargaining position for the later formation of coalitions. However, restrictions such as an impossibility of immediate cap coordination could change our positive results.

Since the presented probabilities are based on a static cost–benefit model, future studies should use more accurate, path-dependent payoffs, effects of leakage and trade feedbacks, and policy instruments such as tariffs. More importantly, the question of how players may arrive at common levels of far-sightedness, common assessments of mitigation costs and benefits and bargaining power, and common beliefs about the process should be studied. Nevertheless, our results seem to justify more hope that a first-best global cap-and-trade system evolves under the Paris Agreement bottom-up, with ambitions increasing over time even if there are at present only few coordinated carbon markets.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41558-018-0079-z.

Received: 6 February 2017; Accepted: 16 January 2018;
Published online: 26 February 2018

References

1. *Adoption of the Paris Agreement* FCCC/CP/2015/L.9/Rev.1 (UNFCCC, 2015); http://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf
2. Green, J. F., Sterner, T. & Wagner, G. A balance of bottom-up and top-down in linking climate policies. *Nat. Clim. Change* **4**, 1064–1067 (2014).
3. Heitzig, J., Lessmann, K. & Zou, Y. Self-enforcing strategies to deter free-riding in the climate change mitigation game and other repeated public good games. *Proc. Natl Acad. Sci. USA* **108**, 15739–15744 (2011).
4. Carraro, C. & Siniscalco, D. Strategies for the international protection of the environment. *J. Public Econ.* **52**, 309–328 (1993).
5. Barrett, S. Self-enforcing international environmental agreements. *Oxford Econ. Pap.* **46**, 878–894 (1994).
6. Finus, M. in *Environmental Policy in an International Perspective* (eds Marsiliani, L., Rauscher, M. & Withagen, C.) 19–49 (Kluwer, Dordrecht, Holland, 2003).
7. Helm, C. International emissions trading with endogenous allowance choices. *J. Public Econ.* **87**, 2737–2747 (2003).
8. Carbone, J. C., Helm, C. & Rutherford, T. F. The case for international emission trade in the absence of cooperative climate policy. *J. Environ. Econ. Manag.* **58**, 266–280 (2009).
9. Smead, R., Sandler, R. L., Forber, P. & Basl, J. A bargaining game analysis of international climate negotiations. *Nat. Clim. Change* **4**, 442–445 (2014).
10. Verendel, V., Johansson, D. J. A. & Lindgren, K. Strategic reasoning and bargaining in catastrophic climate change games. *Nat. Clim. Change* **6**, 6–10 (2015).
11. Ray, D. & Vohra, R. Equilibrium binding agreements. *J. Econ. Theory* **73**, 30–78 (1997).
12. Ray, D. & Vohra, R. A theory of endogenous coalition structures. *Games Econ. Behav.* **26**, 286–336 (1999).
13. Konishi, H. & Ray, D. Coalition formation as a dynamic process. *J. Econ. Theory* **110**, 1–41 (2003).

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility, J. Heitzig and U. Kornek, doi: 10.1038/s41558-018-0079-z, Copyright 2018, The Author(s).

14. de Zeeuw, A. Dynamic effects on the stability of international environmental agreements. *J. Environ. Econ. Manag.* **55**, 163–174 (2008).
15. Biancardi, M. & Villani, G. Largest consistent set in international environmental agreements. *Comput. Econ.* **38**, 407–423 (2011).
16. Osmani, D. *A Note on Computational Aspects of Far-sighted Coalitional Stability.* Hamburg University, Sustainability and Global Change Research Unit Working Papers FNU–176 1–18 (2011).
17. Godal, O. & Holtsmark, B. On the efficiency gains of emissions trading when climate deals are non-cooperative. *Bergen Inst. Res. Econ. Bus. Adm. Work. Pap.* **17**, 1–24 (2011).
18. Flachsland, C., Marschinski, R. & Edenhofer, O. To link or not to link: benefits and disadvantages of linking cap-and-trade systems. *Clim. Policy* **9**, 358–372 (2009).
19. Tuerk, A., Mehling, M., Flachsland, C. & Sterk, W. Linking carbon markets: concepts, case studies and pathways. *Clim. Policy* **9**, 341–357 (2009).
20. Jaffe, J. & Stavins, R. N. *Linkage of Tradable Permit Systems in International Climate Policy Architecture.* NBER Working Paper 14432 (2008).
21. Flachsland, C., Marschinski, R. & Edenhofer, O. Global trading versus linking: Architectures for international emissions trading. *Energy Policy* **37**, 1637–1647 (2009).
22. Ranson, M. & Stavins, R. N. Linkage of greenhouse gas emissions trading systems: learning from experience. *Clim. Policy* **16**, 284–300 (2016).
23. Ellerman, A. D. & Decaux, A. *Analysis of Post-Kyoto CO_2 Emissions Trading Using Marginal Abatement Curves.* MIT Joint Program on the Science and Policy of Global Change Report 40 (1998).
24. Finus, M., van Ierland, E. & Dellink, R. Stability of climate coalitions in a cartel formation game. *Econ. Gov.* **7**, 271–291 (2006).
25. Kalai, E. Nonsymmetric Nash solutions and replications of 2-person bargaining. *Int. J. Game Theory* **6**, 129–133 (1977).
26. Barrett, S. in *Conflicts and Cooperation in Managing Environmental Resources* (ed. Pethig, R.) 11–37 (Springer: Berlin, Heidelberg, 1992).

Acknowledgements

The authors thank K. Lessmann, R. Marschinski, O. Edenhofer, the Policy Instruments Group and the COPAN Flagship Project at the Potsdam Institute for Climate Impact Research for many intense discussions; B. Holtsmark, M. Greaker, C. Hagem, R. Schmidt, and the CREW project for inspiring work; and R. Vohra, P. Menck, N. Marwan, J. Donges and C.-F. Schleussner for helpful comments.

Author contributions

J.H. developed the model and conducted the numerical experiments. Both authors interpreted the study, and wrote and edited the text.

Competing interests

The authors have no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41558-018-0079-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to J.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims inpublished maps and institutional affiliations.

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility, J. Heitzig and U. Kornek, doi: 10.1038/s41558-018-0079-z, Copyright 2018, The Author(s).

Methods

Model overview. As players we consider either two or three hypothetical countries or the six major GHG emitters C(hina), E(urope), F(ormer Soviet Union), I(ndia), J(apan), U(SA).

In each period, the market structure code (MSC) specifies which markets exists (separated by commas), whether there was immediate cap coordination upon market formation (indicated by square brackets), which top-level cap-coordinating coalitions exist in each market (separated by dashes), and any subcoalitions of these (in round brackets). For example, the code [CU],(EF)J-I has two markets: a joint one in C + U with immediate cap coordination, and another in which I sets its caps independently but E + F coordinated their caps before coordinating with J.

Each change to the MSC is called a transition and can be brought about by a move of some set of initiating players which are considered the 'relevant' players for this transition. Players propose, amend, and accept or reject moves based on payoff expectations (called evaluations) and several forms of individual and collective rationality similar to ref. [13], with probabilities depending on a given distribution of bargaining power. Payoff expectations are based on a static payoff function stating each player's payoff in each MSC, on players' beliefs about further changes in MSC, and on their degree of far-sightedness (combining time preferences and trust in the process).

Feasible moves include the linking of markets with or without immediate cap coordination, forming and merging of cap-coordinating coalitions, and optionally unanimous or unilateral termination of links or mergers ('reversibility').

Assumed forms of rationality are: accepting only profitable move proposals, 'amending' (that is, changing) move proposals that are dominated by a more profitable move of some subgroup[13], proposing only favourite undominated profitable moves, and collectively forming correct beliefs about the process.

The model output is a process diagram specifying a probability for each feasible transition between MSCs. Because of rationality, these probabilities depend on payoff expectations by player and MSC, as well as an assumed bargaining power distribution. Since players form correct beliefs, payoff expectations must equal discounted average long-term payoffs ('evaluations'), which in turn depend on (the believed) move probabilities. The resulting system of nonlinear equations between probabilities and expected payoffs is solved numerically, resulting in an equilibrium process that represents a consistent combination of transition probabilities based on rationality and correct beliefs about expected payoffs. In other words, an equilibrium process is a set of commonly believed ('subjective') transition probabilities on the basis of which all players' rational behaviour would bring about these very same probabilities in reality (that is, as 'objective' probabilities). In short, an equilibrium process is a set of common beliefs that is consistent with the common assumption of rational behaviour. One can prove that for all parameter settings, there is at least one equilibrium process, and sometimes there are many. Exactly as for proving the existence of many other game-theoretic equilibria (for example, Nash equilibrium), that proof consists in applying Kakutani's fixed point theorem to a suitable set-valued function (here the one that relates subjective to objective transition probabilities, see Supplementary Section 1.1.2).

See the following section 'General game structure' and Supplementary Section 1.1 for details.

General game structure. *Players and notation for carbon market structure.* We assume a set P of $N > 0$ players. In each period, each player i (that is, a central authority in the respective country or world region; for example, the government) first chooses their domestic emissions cap c_i individually, issuing that many permits to their domestic industry or population. These can then be traded freely in a domestic or international carbon market such as the European Union's Emissions Trading System (EU ETS). In the terminology of ref. [21], this means that we consider a 'bottom-up' cap-and-trade architecture in which companies or households are trading permits in a sufficiently 'integrated' international market at a market-wide equalized price, while governments only issue permits but do not trade them directly, instead of a 'top-down' architecture in which governments trade permits directly (as in the Kyoto Protocol). For simplicity, if several carbon markets have been linked, we treat them as one large market and do not analyse the trade in its parts individually while they are linked. This corresponds two 'two-way direct links' in the terminology of ref. [20], also known as 'formally linked' markets in the terminology of ref. [21].

We represent the market structure by a code in which the markets are separated by commas and the members by dashes. For example, the code C-U,E-F-J,I represents three markets, a domestic one consisting of player I, one international with members E, F and J, and one international with members C and U. After trading, player i's actual emissions $e_i(t)$ equal its post-trade amount of permits, so that

$$\sum_{i\in P} e_i = \sum_{i\in P} c_i =: E \tag{3}$$

and she gets a pre-transfer payoff of $f(e_j : j \in P)$ depending on everyone's actual emissions via some function f to be specified later. A player's static payoffs $\pi_i(x)$ are then the sum of $f(e_j : j \in P)$ and any transfers by which the coalitions that i is a member of implement their surplus sharing (see the section: Derivation of static payoffs).

Notation for cap-coordinating coalitions. Within each market, players might be organized in a tree-like hierarchy of coalitions as in ref. [27]. A coalition in our sense is a subset K of the members of a market M that agree to coordinate their cap choices in some way. Such an agreement might have been signed by individual players or by sub-coalitions that have formed earlier. We assume that cap choices are coordinated in such a way that the surplus (the difference between the post- and pre-agreement coalitional payoffs) is distributed in some fixed proportions given by the bargaining power of the signatories (see below). We treat individual players as one-member 'coalitions'. There is no explicit cap coordination between the top-level coalitions in a market.

We represent the coalition hierarchy in a market by a code in which the top-level coalitions are separated by dashes and the lower-level coalitions are identified by parentheses. For example, the code EF-J represents a market with members E, F and J, in which E and F have formed a coalition by agreeing to coordinate their cap choices, while J chooses its caps individually. If the coalition EF in a later period signs a further agreement with J, the code becomes (EF)J. If all three had agreed immediately without a preceding bilateral agreement, we would write EFJ instead. Note that because of the assumed proportional surplus-sharing rule (see the section: Derivation of static payoffs), while the total cap choice of E, F and J will be the same in these two situations, they will share this total cap in different ways in the two situations, since the pre-agreement payoffs are those in EF-J in the first situation but those in E-F-J in the second. Hence payoffs depend on both top-level and lower-level coalition structure, and it is important to distinguish the cases (EF)J and EFJ.

Market linkage and notation for states and moves. Markets can be linked in two ways: Either several markets such as C-U and EF-J are linked without immediate coordination of caps, thus becoming a new larger market C-EF-J-U, or several markets that have already reached full internal cap coordination, such as CU and (EF)J, are linked with immediate overarching cap coordination, which is then indicated by square brackets: [(CU)((EF)J)]. Once a market is formed by the second type of agreement, that is, with immediate cap coordination, it is assumed that it can no longer be linked with further markets by the first kind of agreement, that is, without immediate further cap coordination. In other words, the markets [(CU)((EF)J)] and I can only be linked to form [[(CU)((EF)J)]I], while [(CU)((EF)J)]-I is impossible. Of course, the markets CU and (EF)J could also develop into CU-(EF)J, then into (CU)((EF)J) in a second step, and then into (CU)((EF)J)-I. But although (CU)((EF)J) and [(CU)((EF)J)] will get the same joint payoff, their cap distributions will differ, again because of the surplus-sharing rule which compares the payoff in (CU)((EF)J) with that in the one market CU-(EF)J but compares the payoff in [(CU)((EF)J)] with that in the two markets CU,(EF)J instead to determine surpluses.

Combining the market structure and coalition hierarchy codes to state codes and indicating moves between states with arrows labelled by the subset of players who are required for initiating that move, the above fictitious example process would be denoted

$$\begin{aligned}
& \text{C–U,E–F–J,I} \\
\xrightarrow{\text{EF}}\ & \text{C–U,EF–J,I} \\
\xrightarrow{\text{EFJ}}\ & \text{C–U,(EF)J,I} \\
\xrightarrow{\text{CU}}\ & \text{CU,(EF)J,I} \\
\xrightarrow{\text{CEFJU}}\ & \text{[(CU)((EF)J)],I} \\
\xrightarrow{\text{CEFIJU}}\ & \text{[[(CU)((EF)J)]I]}
\end{aligned}$$

The number of theoretically possible states grows faster than exponentially in the number of players. For five or six players, the model has already 2729 or 41,106 states, respectively; hence we restrict our analysis to six players at this time. Fortunately, our results verify the intuition that only a very small number of these possible states occur with positive probability. The actual process might then, for example, look as depicted in Figs. 2,3 and 4 where the arrows are labelled with transition probabilities and those players that favour the move.

Individual and collective rationality, far-sightedness. In order to decide which moves to consider, we assume that players apply certain principles of individual and collective rationality, trying to influence the market structure and coalition hierarchy to optimize their average, properly discounted long-term payoffs, which we call their evaluations, denoted ℓ_i. We assume that they do so in a far-sighted way, anticipating the further development of the structure. We model the level of this far-sightedness via a number $\delta \in (0,1)$ used in the discounting of prospective future states' static payoffs π_i. This far-sightedness parameter δ can be interpreted as a combined measure of time discounting, period length, and trust in the process (see below for details).

In contrast to some other game-theoretic models of coalition formation, we do not assume that the changes to the market structure and coalition hierarchy follow a specific bargaining protocol precisely prescribing who can propose which move at what time to whom, since in the climate context negotiations are probably not following such restrictive rules. Instead, we assume that in each period, the set of

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility, J. Heitzig and U. Kornek, doi: 10.1038/s41558-018-0079-z, Copyright 2018, The Author(s).

initiators of any feasible move can consider its realization if they all agree to do so. If several different moves are considered in a period, however, it depends on other factors than only rationality principles which move will actually get realized. In the model this is represented by assigning probabilities to moves on the basis of all players' preferences and on assumptions about their bargaining power.

We consider different levels of rationality. In the weakest case, any move might be considered that is individually profitable for each of its initiators, using one of several concepts of profitability to be discussed below. On the medium level of rationality, only those profitable moves might be considered which are undominated in the sense that its initiators cannot initiate a different move which they all prefer (this corresponds to the approach in ref. [13]). Even stronger, we assume that an undominated move will only be considered if it is the favourite undominated move of at least one player, be it an initiator of the move or not, based on the assumption that no international agreement will come about without at least one country pressing for its realization. We could have considered an even higher level of collective rationality in which players can find a consensus move which no player favours but which all players prefer to the otherwise resulting lottery of favourite moves, as in ref. [28]. With the long-term profitability concept of our model, however, such consensus moves are automatically identified as the only profitable moves in an equilibrium process.

The remaining uncertainty about which move will actually be realized is then expressed as a probability distribution over the thus determined set of considered moves, assuming that only one of them will be realized in each period even when there are several moves considered by disjoint sets of initiators which could in principle be realized at the same time. The latter assumption is justified by the fact that usually a move by one set of players also affects the payoffs of other players, so that when a certain move is about to be made by some of the major emitters, it seems plausible to assume that the other players will wait with their attempt of an additional move until it becomes clear whether the first move will actually be realized.

Derivation of static payoffs. *Assumptions.* For the six-player case, we use an analytically derived form of the payoff function π_i that results from the following assumptions:

- Abatement costs are cubic functions of actual domestic abatement.
- Abatement benefits are linear functions of global abatement.
- Emissions trading equalizes the price with all marginal abatement costs.
- Before the trading, all top-level coalitions simultaneously choose their coalitional caps to maximize their respective joint payoffs, anticipating its effect on trading (that is, on traded amounts and price), leading to a global Nash equilibrium between all top-level coalitions of all markets.
- Each coalition allocates their coalitional cap to its members so that the surplus is shared in some exogenously given fixed proportions.

The functional form and coefficients of the abatement cost and benefit functions for the six real-world emitters are taken at this point from the STACO model (ref. [24] version) which calibrates its benefit estimates to the vastly used DICE model of ref. [29] and takes its cost estimates from ref. [23], because that model presents a good trade-off between tractability and qualitative real-world relevance. For future applications, one may use newer estimates, for example, derived from ref. [30] or from more sophisticated models such as the one in ref. [8]. To keep our numbers comparable to those in ref. [24], we report e_i in metric gigatons of carbon (GtC) per 100 years and π_i in US$ billion per 100 years.

Derivation of coalitional payoffs. Given the actual emissions e_i, the STACO model expresses individual payoff in terms of individual abatement contributions $q_i = e_i^0 - e_i > 0$ with respect to some fixed reference ('business as usual') emissions e_i^0 since this formulation makes it easier to compare the abatement game with other public good games. In the linearized static version of STACO that we use here, benefits from global abatement (avoided damages from climate change) are a linear function 37.4 US$ per tC × σ_i /1000 × Q of global contributions $Q = \sum_{i \in P} q_i = E^0 - E$, and costs of abatement are a cubic function

$$g_i(q_i) = \frac{a_i q_i^3}{3} + \frac{b_i q_i^2}{2} \tag{4}$$

of individual contributions, where the coefficients σ_i, a_i, b_i are given in Supplementary Table 1 using calibration I from ref. [24]. Together with the emissions trade balance, individual payoffs of a member i of a market M in terms of caps and emissions are then

$$\pi_i = \sigma_i(E^0 - E) - g_i(e_i^0 - e_i) + p_M(c_i - e_i) \tag{5}$$

where p_M is the market price in M.

The remaining derivation is a straightforward application of the one in ref. [7] to the case of several markets. We assume that each emissions market M has perfect competition, so that the marginal abatement costs at the post-trade abatement levels are equal to the market price for all market members,

$$g_i'(e_i^0 - e_i) = p_M \tag{6}$$

for all $i \in M$ (see Supplementary Fig. S1) for the corresponding marginal abatement cost curves). Since the market's cap equals the market's emissions,

$$\begin{aligned} c_M &= \sum_{i \in M} c_i \\ &= e_M = \sum_{i \in M} e_i = \sum_{i \in M} [e_i^0 - (g_i')^{-1}(p_M)], \end{aligned} \tag{7}$$

the price p_M can be seen as a function of c_M whose derivative is related to individual emissions via the theorem on implicit functions as

$$\frac{\mathrm{d}}{\mathrm{d}c_M} p_M = -1 / \sum_{i \in M} \frac{1}{g_i''(e_i^0 - e_i)} < 0. \tag{8}$$

Now we assume that each top-level coalition K in M acts as an output cartel that chooses its cap $c_K = \sum_{i \in K} c_i$ to maximize its joint payoffs,

$$\begin{aligned} \pi_K &= \sigma_K Q - \sum_{i \in K} g_i(e_i^0 - e_i) + p_M(c_K - e_K) \\ &= \sigma_K Q - \sum_{i \in K} [g_i(e_i^0 - e_i) + p_M e_i] + p_M c_K, \end{aligned} \tag{9}$$

taking the caps $c_{K'}$ of all other top-level coalitions $K' \neq K$ as given, where σ_K, e_K are the coalitional aggregates of σ_i, e_i. The corresponding first-order condition is

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}c_K} \pi_K \\ &= \sigma_K \frac{\mathrm{d}}{\mathrm{d}c_K} Q + \sum_{i \in K} [g_i'(e_i^0 - e_i) - p_M] \frac{\mathrm{d}}{\mathrm{d}c_K} e_i + p_M + (c_K - e_K) \frac{\mathrm{d}}{\mathrm{d}c_K} p_M \\ &= p_M - \sigma_K + (c_K - e_K) \frac{\mathrm{d}}{\mathrm{d}c_M} p_M \end{aligned} \tag{10}$$

by Eq. 6, where the last term reflects the fact that the coalition is not a 'price-taker' but is aware of its choice's effect on the price. If there are n_M top-level coalitions in M, their simultaneous optimization leads to a unique Nash equilibrium which can easily be found analytically by summing the above condition over all n_M coalitions, giving

$$0 = n_M\, p_M - \sigma_M + (c_M - e_M) \frac{\mathrm{d}}{\mathrm{d}c_M} p_M = n_M\, p_M - \sigma_M \tag{11}$$

by Eq. 7. Hence the market price is simply

$$p_M = \sigma_M / n_M \tag{12}$$

actual individual emissions are

$$e_i = e_i^0 - (g_i')^{-1}(p_M) = e_i^0 - \frac{\sqrt{b_i^2 + 4 a_i p_M} - b_i}{2 a_i} \tag{13}$$

by Eq. 6, the coalition's cap choice is

$$c_K = e_K + (p_M - \sigma_K) \sum_{i \in M} \frac{1}{2 a_i (e_i^0 - e_i)} \tag{14}$$

by Eqs. 8,10, and 13, and all coalitions' payoffs are given by Eq. 9.

From this general payoff structure, ref. [7] derives several effects of establishing a global carbon market without cap coordination that translate into our setting as follows:

- A coalition K in a market M is a permit seller iff $\sigma_K < p_M$ (follows from Eq. 10).
- When markets are linked without coordinating caps further than before, permit sellers might increase their caps and global emissions might actually increase instead of decrease.
- Independently of whether such a linkage decreases or increases the market's cap, it might or might not be profitable for all members.

At first glance, all this might indicate that the immediate coordination of caps when linking markets is the preferable option since it surely gives a positive surplus that can be distributed via cap redistribution to make sure that all members profit from it. Such myopic reasoning, however, neglects the possibility that also after a linkage without cap coordination, caps might later on be coordinated, and some coalitions might prefer such a two-step process since its first step puts them in a more comfortable bargaining situation for the second step. It is precisely such effects and the resulting conflicts that our dynamic model uncovers.

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility, J. Heitzig and U. Kornek, doi: 10.1038/s41558-018-0079-z, Copyright 2018, The Author(s).

Surplus-sharing and bargaining power. Finally, each top-level coalition K determines their surplus payoff $\Delta\pi_K = \pi_K - \pi_K^0$ by comparing their joint payoff π_K with the joint payoff $\pi_K^0 = \sum_{i \in K} \pi_i^0$ their members i would get in the following reference state: remove coalition K from the coalition hierarchy, and if K is of the immediate-coordination form […], also split the corresponding market into one market for each of the resulting top-level coalitions. For example, for K = (EJ) U in state C-(EJ)U,FI the reference state is C-EJ-U,FI, while for K = [C(EJ)U] in state [C(EJ)U],FI the reference state is C,EJ,FI,U. Then coalition K allocates their joint cap c_K in such a way that each player $i \in K$ gets a share of this surplus that is proportional to their bargaining power w_i, so that

$$\pi_i = \pi_i^0 + \frac{\Delta\pi_K w_i}{\sum_{j \in K} w_j} \tag{15}$$

A possible interpretation of this surplus-sharing rule that relates it to traditional solution concepts of cooperative game theory is this: each player gets its weighted Shapley value or, equivalently, its share as determined by the asymmetric Nash bargaining solution[25], in the unanimity game v with $v(K') = \Delta\pi_K$ if $K' \supseteq K$ and $v(K')=0$ otherwise, using the weights w_i (compare ref. [31] which also discuss using population as weight). The underlying rationale is that the reference state is the only alternative state that could realistically be reached on short notice, by terminating only one top-level agreement, so that the value of each player's outside option is simply its payoffs in that reference state.

For player's bargaining power weights, we use a subjectively chosen distribution that aims at a simple compromise between the following possible choices (see Supplementary Table 1):

- w_i=population of i.
- w_i=GDP of i in US$.
- w_i=σ_i (climate "vulnerability").
- w_i=1 (equal bargaining power).
- An 'egalitarian' approach that leads to equal per-capita surplus in purchasing power parities (PPP):

$$\begin{aligned} w_i = \ & (\text{population of } i) \\ & \times (\text{PPP in currency of } i) \\ & \times (\text{exchange rate from } i \text{ to US\$}) \end{aligned}$$

Total payoff uncertainty. To assess the stochasticity of the process, we use the metric $\sqrt{\sum_i \mathrm{Var}(L_i)}$ where $L_i(0)$ is player i's actual discounted long-term payoff when starting at the root node, and the variance is over the different realizations of the actual path towards cooperation that make L_i a random variable. Like its expected value $\ell_i = \mathrm{E}(L_i)$ (Eq. 1), L_i can be calculated recursively,

$$L_i(X(t)) = (1-\delta)\pi_i(X(t)) + \delta L_i(X(t+1))$$

where the random variable $X(t)$ is the state in period t.

Data availability. The authors declare that the data supporting the findings of this study are available within the article and its Supplementary Information file. Additional model output is available on request from the first author.

References

27. Heitzig, J. Efficiency in face of externalities when binding hierarchical agreements are possible. *Game Theory Bargain. Theory eJournal* **3**, 1–16 (2011).
28. Heitzig, J. & Simmons, F. W. Some chance for consensus: Voting methods for which consensus is an equilibrium. *Social. Choice Welf.* **38**, 43–57 (2012).
29. Nordhaus, W. D. *Managing the Global Commons: the Economics of Climate Change* (MIT Press, Cambridge, MA, 1994).
30. Nordhaus, W. D. Economic aspects of global warming in a post-Copenhagen environment. *Proc. Natl Acad. Sci. USA* **107**, 11721–11726 (2010).
31. Kalai, E. & Samet, D. On weighted Shapley values. *Int. J. Game Theory* **16**, 205–222 (1987).

Reprinted by permission from Springer Nature: Springer Nature, Nature Climate Change, Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility, J. Heitzig and U. Kornek, doi: 10.1038/s41558-018-0079-z, Copyright 2018, The Author(s).

4.3 *Earth system analysis and planetary boundary interactions*

This last section features publications which focus on coevolutionary interactions of societal management and Earth system dynamics within the planetary boundaries.

Due to the increasing levels of atmospheric carbon resulting from anthropogenic fossil fuel burning and land-use change, several climate engineering methods like terrestrial carbon dioxide removal (tCDR) have recently been discussed. In "Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal" [Heck et al., 2016] we analyzed the coevolutionary interaction of societal interventions via tCDR and the natural dynamics of the Earth's carbon cycle. Our study elaborated the danger of transgressing certain planetary boundaries when applying tCDR in a business-as-usual scenario.

Deforestation in the Amazon has enormous consequences for the ecosystem and the climate. The potentials of management options such as intensifying cattle ranching to reduce deforestation are controversial. In "Can intensification of cattle ranching reduce deforestation in the Amazon? Insights from an agent-based social-ecological model" [Müller-Hansen et al., 2019], we examined the social-ecological interplay using a multi-agent adaptive network model that links social learning and ecological processes with market dynamics.

Earth Syst. Dynam., 7, 783–796, 2016
www.earth-syst-dynam.net/7/783/2016/
doi:10.5194/esd-7-783-2016

Earth System Dynamics
Open Access

Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal

Vera Heck[1,3], **Jonathan F. Donges**[1,2], **and Wolfgang Lucht**[1,3,4]

[1]Earth System Analysis, Potsdam Institute for Climate Impact Research, Telegraphenberg A62, 14473 Potsdam, Germany
[2]Stockholm Resilience Centre, Stockholm University, Kräftriket 2B, 114 19 Stockholm, Sweden
[3]Department of Geography, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany
[4]Integrative Research Institute on Transformations of Human-Environment Systems, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany

Correspondence to: Vera Heck (heck@pik-potsdam.de)

Received: 4 May 2016 – Published in Earth Syst. Dynam. Discuss.: 17 May 2016
Revised: 26 September 2016 – Accepted: 27 September 2016 – Published: 31 October 2016

Abstract. The planetary boundaries framework provides guidelines for defining thresholds in environmental variables. Their transgression is likely to result in a shift in Earth system functioning away from the relatively stable Holocene state. As the climate system is approaching critical thresholds of atmospheric carbon, several climate engineering methods are discussed, aiming at a reduction of atmospheric carbon concentrations to control the Earth's energy balance. Terrestrial carbon dioxide removal (tCDR) via afforestation or bioenergy production with carbon capture and storage are part of most climate change mitigation scenarios that limit global warming to less than 2 °C.

We analyse the co-evolutionary interaction of societal interventions via tCDR and the natural dynamics of the Earth's carbon cycle. Applying a conceptual modelling framework, we analyse how the degree of anticipation of the climate problem and the intensity of tCDR efforts with the aim of staying within a "safe" level of global warming might influence the state of the Earth system with respect to other carbon-related planetary boundaries.

Within the scope of our approach, we show that societal management of atmospheric carbon via tCDR can lead to a collateral transgression of the planetary boundary of land system change. Our analysis indicates that the opportunities to remain in a desirable region within carbon-related planetary boundaries only exist for a small range of anticipation levels and depend critically on the underlying emission pathway. While tCDR has the potential to ensure the Earth system's persistence within a carbon-safe operating space under low-emission pathways, it is unlikely to succeed in a business-as-usual scenario.

1 Introduction

Rockström et al. (2009) introduced the concept of a safe operating space (SOS) for humanity, delineated by nine global planetary boundaries, some of which take into account the existence of tipping points or nonlinear thresholds in the Earth system (Lenton et al., 2008; Schellnhuber, 2009; Kriegler et al., 2009) and may frame sustainable development. Particularly, the state of the Earth system with respect to climate change has received strong political attention as atmospheric carbon concentrations have already entered the uncertainty zone of the planetary boundary of climate change, set at an atmospheric CO_2 concentration of 350 to 450 ppmv (Steffen et al., 2015).

The Paris climate agreement (UNFCCC, 2015) aims at limiting global temperature increase to well below 2 °C above pre-industrial levels, while greenhouse gas emissions are still currently growing. Fuss et al. (2014) have highlighted that more than 85 % of IPCC scenarios that are consistent with the 2 °C goal require net negative emis-

Published by Copernicus Publications on behalf of the European Geosciences Union.

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

sions before 2100. Particularly, terrestrial carbon dioxide removal (tCDR) via afforestation or large-scale cultivation of biomass plantations for the purpose of bioenergy production has been included in recent IPCC scenarios (van Vuuren et al., 2011; Kirtman et al., 2013). Furthermore, tCDR has been proposed as a climate engineering (CE) method that could be applied in case global efforts in mitigating anthropogenic greenhouse gas emissions fail to prevent dangerous climate change (Caldeira and Keith, 2010).

In the context of the SOS framework, tCDR via large-scale biomass plantations could extract carbon from the atmosphere via the natural process of photosynthesis (Shepherd et al., 2009). If the carbon accumulated in biomass is harvested and stored in deep reservoirs or used for bioenergy production in combination with carbon capture and storage (Caldeira et al., 2013), further transgression of the climate change boundary and initial transgression of the ocean acidification boundary could be prevented. On the other hand, tCDR is likely to have unintended impacts on other Earth system components besides atmospheric carbon concentrations that is mediated by the global cycles of carbon, water and other biogeochemical compounds (Vaughan and Lenton, 2011). For example, large-scale biomass plantations would require substantial amounts of fertiliser, irrigation water and land area, driving the Earth system closer to the planetary boundaries for biogeochemical flows, freshwater use and land system change, respectively (Heck et al., 2016). The tCDR in the form of afforestation would not be accompanied by most of these negative trade-offs. However, afforestation only has a limited potential to increase the terrestrial carbon storage while all emitted fossil carbon remains a part of the active carbon cycle. Thus, the potentials of tCDR via afforestation are small and afforestation is not included as a tCDR method in this study.

Social and political actions are important drivers of tCDR. The willingness to engage in CE or mitigation is based on monitoring of the climate system and can be expected to increase as the climate system approaches the normatively assigned climate change boundary. A holistic assessment and systemic understanding of CE therefore requires an analysis of the social and ecological co-evolutionary system.

A dynamic integration of complex interactions between the social and ecological components of the Earth system to simulate in detail the co-evolution of societies and the environment is currently unfeasible due to fundamental conceptual problems and high computational demands on both modelling sides (van Vuuren et al., 2012, 2016). An emerging field of low-complexity models explores new pathways for understanding social–ecological Earth system dynamics (e.g. Brander and Taylor, 1998; Kellie-Smith and Cox, 2011; Jarvis et al., 2012; Anderies et al., 2013; Motesharrei et al., 2014). For example, first simulation approaches have been reported using such conceptual models to simulate the interaction between human climate monitoring and societal action in the form of transitions to renewable energy (Jarvis et al., 2012) or climate engineering (MacMartin et al., 2013). While not aiming for realism in their quantitative evaluations, the low complexity of such conceptual models allows to understand the structure and effects of dominating feedbacks and their leading interactions, which are otherwise often hidden in the complexity of state-of-the-art full-complexity Earth system models.

In this paper, we provide a conceptual but systematic analysis of the nonlinear system response to using tCDR for steering the Earth system within the SOS defined by planetary boundaries as quantified by Rockström et al. (2009) and Steffen et al. (2015). Specifically, we analyse how the trade-offs between tCDR and other planetary boundaries depend on the achievable rate and threshold of tCDR implementation; and whether particular combinations of climate and management parameterisations can safeguard a persistence within the SOS. As a starting point, we focus on a subset of the nine proposed planetary boundaries that are most important in the context of tCDR. These are the carbon-related boundaries on climate change, ocean acidification and land system change.

We utilise a conceptual model of the carbon cycle and expand it to explore feedbacks within and between societal and ecological spheres, while being sufficiently simple to permit an analysis of its state and parameter spaces in the form of constrained stability analysis similar to van Kan et al. (2016). We do not aim to provide a quantitative assessment because in this exploratory study, we choose to use a computationally efficient conceptual model to shed light onto the qualitative structure of co-evolutionary dynamics. The approach proposed here can be transferred to models of higher complexity to the extent that this is computationally feasible.

This paper is structured as follows: following the introduction (Sect. 1) we present a co-evolutionary model of societal monitoring and tCDR intervention in the Earth's carbon cycle and related parameter calibration procedures (Sect. 2). Subsequently, we present and discuss our results (Sect. 3) and finish with conclusions (Sect. 4).

2 Methods

In social–ecological systems modelling, societal influences and ecological responses are recognised as equally important (Berkes et al., 2000). Therefore, it can be considered essential that representations of social and ecological systems are of the same order of complexity. Increasing complexity of only one model component would not increase the accuracy of information generated by the full coupled model, but would greatly increase computational demand. In view of our objective, we require a sufficiently simple model that conceptually captures the most important processes of global carbon dynamics with respect to planetary boundaries, as well as a stylised societal management feedback loop consisting of tCDR interventions and monitoring of the climate system.

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

2.1 Co-evolutionary model of societal monitoring and tCDR intervention in the carbon cycle

The basis of our co-evolutionary model is the conceptual carbon cycle model by Anderies et al. (2013). The model covers the most basic interactions between terrestrial, atmospheric and marine carbon pools, and was developed specifically to enable a bifurcation analysis of carbon-related planetary boundaries and their interactions. We modified atmosphere–land interactions for a better representation of empirically observed Earth system carbon dynamics and extended the model by a stylised societal management feedback loop mimicking the current focus of international policy processes on climate change. We calibrated the model in order to represent global carbon cycle dynamics consistent with observational data and simulations from detailed high-resolution Earth system models (Sect. 2.2). In the following, we provide an overview of the fundamental model equations. A detailed motivation of the model design and underlying assumptions are given in Anderies et al. (2013).

The adapted model consists of five interacting carbon pools: land $C_\mathrm{t}\ (t)$, atmosphere $C_\mathrm{a}(t)$, upper-ocean $C_\mathrm{m}\ (t)$, geological fossil reservoirs $C_\mathrm{f}(t)$ and a potential CE carbon sink $C_\mathrm{CE}(t)$ (Fig. 1). All model equations are summarised in Table 1. Note that only the upper-ocean carbon pool is included because the movement of carbon into the deep ocean occurs on longer timescales relative to those of interest, as discussed by Anderies et al. (2013). The land carbon pool combines soil and vegetation carbon pools, implying a simple proportional partitioning of aboveground and belowground carbon pools (Anderies et al., 2013). These simplifications have been adopted because they reduce the number of state variables and we were able to qualitatively reproduce the dynamics of observed carbon pool evolution with the adapted model.

The co-evolutionary dynamics of the system is determined by Eqs. (1)–(5). Conservation of mass (Eq. 1) dictates that the active carbon in the system, i.e. the sum of terrestrial, atmospheric and maritime carbon is equal to the active carbon at pre-industrial times (C_0) plus carbon released from fossil reservoirs ($C_\mathrm{r}(t)$) minus carbon extracted via tCDR ($C_\mathrm{CE}(t)$) to permanent stores. Fossil carbon release (Eq. 2) is approximated by a logistic function parameterised by the maximum emitted carbon c_max and rate of carbon release r_i.

The social management feedback loop is motivated by proposals of CE as a management intervention in response to intolerable levels of global warming. It comprises atmospheric carbon monitoring and tCDR action conditional on the proximity to a critical threshold of atmospheric carbon content (Eq. 3). CE action is implemented via a tCDR carbon offtake from terrestrial carbon ($H_\mathrm{CE}(t)$) and storage in a permanent (geological) sink C_CE. Carbon offtake for tCDR (Eq. 11) is defined analogous to human offtake for agriculture or land-use change (Eq. 13), however, with a dynamic offtake rate $\alpha_\mathrm{CE}(C_\mathrm{a}(t))$ (Eq. 12).

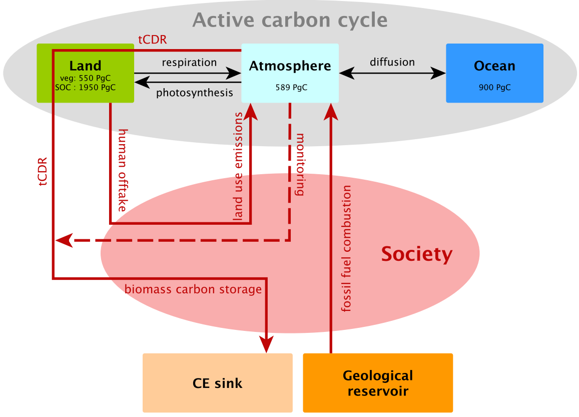


Figure 1. Structure of the co-evolutionary model of societal monitoring and terrestrial carbon dioxide removal (tCDR) intervention in the carbon cycle including simulated components of the carbon cycle as well as a societal management feedback loop and their interactions. Carbon fluxes are indicated as solid lines and coloured red if influenced by society. Carbon values in the boxes indicate estimates of pre-industrial carbon pools in the year 1750 AD (Batjes, 1996; Ciais et al., 2013). CE sink is the climate engineering sink.

The tCDR characteristics are governed by three parameters: (i) implementation threshold ($\widetilde{C_\mathrm{a}}$) in terms of atmospheric carbon content, representing societal foresightedness, (ii) maximally achievable rate of tCDR (α_max), a measure of societies' efforts, as well as biogeochemical constraints and (iii) the slope of tCDR implementation (s_CE), parameterising social and economic implementation capacities. Figure 2 depicts an exemplary tCDR trajectory for constant terrestrial carbon in Eq. (11) for two values of s_CE. The implementation time can be computed from the slope of tCDR implementation by using current increase rates of atmospheric carbon as a conversion factor. With current increase rates of approximately 2 ppmv a^{-1} (Tans and Keeling, 2015), the two depicted values of s_CE correspond to tCDR ramp-up times of approximately 20 and 40 years (from 10 to 90 % capacity) for $s_\mathrm{CE} = 0.1\,\mathrm{ppmv}^{-1}$ (solid) and $s_\mathrm{CE} = 0.05\,\mathrm{ppmv}^{-1}$ (dashed), respectively.

The atmosphere–ocean carbon feedback (Eq. 4) is governed by diffusion, which in the model is assumed to depend on the difference between atmospheric and maritime carbon pools.

Land–atmosphere interaction is determined by both ecological and social processes: the net ecosystem productivity (Eq. 6), tCDR offtake (Eq. 11) and other human offtake for agriculture and other land use (Eq. 13), respectively.

Net ecosystem productivity is given by the net carbon flux of photosynthesis (Eq. 8) and respiration (Eq. 9), multiplied by the terrestrial carbon pool and a logistic dampening function which represents competition for space, sunlight, water or nutrients. Both photosynthesis and respiration are con-

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

Table 1. Summary of equations describing the co-evolutionary model of societal monitoring and tCDR intervention in the carbon cycle building upon Anderies et al. (2013). The unit a is years.

Process	Equation	
Conservation of mass	$C_\mathrm{t}(t)+C_\mathrm{a}(t)+C_\mathrm{m}(t)=C_0+C_\mathrm{r}(t)-C_\mathrm{CE}(t)$	(1)
Fossil carbon release	$\dot{C}_\mathrm{r}(t)=r_i C_\mathrm{r}(t)(1-\frac{C_\mathrm{r}(t)}{c_\mathrm{max}})$	(2)
CE carbon storage	$\dot{C}_\mathrm{CE}(t)=H_\mathrm{CE}(C_\mathrm{t}(t),C_\mathrm{a}(t))$	(3)
Atmosphere–ocean diffusion	$\dot{C}_\mathrm{m}(t)=a_\mathrm{m}(C_\mathrm{a}(t)-\beta C_\mathrm{m}(t))$	(4)
Terrestrial carbon flux	$\dot{C}_\mathrm{t}(t)=\mathrm{NEP}(C_\mathrm{a}(t),C_\mathrm{t}(t))-H(C_\mathrm{t}(t))-H_\mathrm{CE}(C_\mathrm{t}(t),C_\mathrm{a}(t))$	(5)
Net ecosystem productivity	$\mathrm{NEP}(C_\mathrm{a}(t),C_\mathrm{t}(t),T(t))=r_\mathrm{tc}\,[P(T(t))-R(T(t))]\,C_\mathrm{t}(t)\left[1-\frac{C_\mathrm{t}(t)}{K(C_\mathrm{a}(t))}\right]$	(6)
Terrestrial carbon carrying capacity	$K(C_\mathrm{a}(t))=a_k e^{-b_k C_\mathrm{a}(t)}+c_k$	(7)
Photosynthesis	$P(T(t))=a_\mathrm{p}T(t)^{b_\mathrm{p}}e^{-c_\mathrm{p}T(t)}$	(8)
Respiration	$R(T(t))=a_\mathrm{r}T(t)^{b_\mathrm{r}}e^{-c_\mathrm{r}T(t)}$	(9)
Temperature	$T(C_\mathrm{a}(t))=a_T C_\mathrm{a}(t)+b_T$	(10)
tCDR offtake flux	$H_\mathrm{CE}(C_\mathrm{t}(t),C_\mathrm{a}(t))=\alpha_\mathrm{CE}(C_\mathrm{a}(t))C_\mathrm{t}(t)$	(11)
Societal tCDR offtake rate	$\alpha_\mathrm{CE}(C_\mathrm{a}(t))=\alpha_\mathrm{max}\left(1+\exp(-s_\mathrm{CE}(C_\mathrm{a}(t)-\widetilde{C_\mathrm{a}})\right)^{-1}$	(12)
Other human biomass offtake flux	$H(C_\mathrm{t}(t))=\alpha C_\mathrm{t}(t)$	(13)

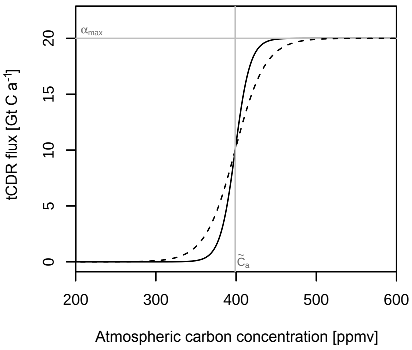


Figure 2. Sigmoidal dependence of the tCDR flux on atmospheric carbon concentrations for two values of the tCDR implementation capacity parameter (slope): $s_\mathrm{CE}=0.1\,\mathrm{ppmv}^{-1}$ (solid line) and $s_\mathrm{CE}=0.05\,\mathrm{ppmv}^{-1}$ (dashed line). The threshold parameter ($\widetilde{C_\mathrm{a}}$) is set at 400 ppmv atmospheric carbon concentration and the potentially achievable tCDR flux is parameterised with $\alpha_\mathrm{max}=20\,\mathrm{Gt\,C\,a}^{-1}$.

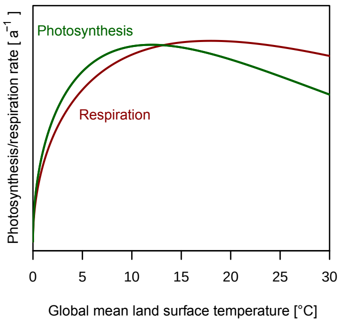


Figure 3. Modelled photosynthesis and respiration rates as a function of global mean land surface temperature.

tinuous functions of global land temperature ($T(t)$, Eq. 10), which in turn depends linearly on atmospheric carbon content. It is important to note that in our model, respiration exceeds photosynthesis for higher temperatures (Fig. 3). The state of equilibrium of the terrestrial carbon pool is thus determined by the land surface temperature, as well as the terrestrial carbon carrying capacity (Eq. 7) in the density function. In contrast to Anderies et al. (2013), we implement a dynamic terrestrial carbon carrying capacity as a function of atmospheric carbon content. This is motivated by a number of factors such as CO_2 fertilisation and a higher water-use efficiency under higher atmospheric carbon concentrations, as well as higher average vegetation density in a warmer world (e.g. Drake et al., 1997; Keenan et al., 2013). For low atmospheric carbon we assume a rapid increase in terrestrial carbon storage capacity as a function of atmospheric carbon concentration and a saturation of storage capacity for high atmospheric carbon, in line with assessments of coupled carbon cycle climate models (Heimann and Reichstein, 2008). The functional relationship in Eq. (7) follows these constraints for chosen parameter values (Sect. 2.2).

2.2 Calibration of model parameters

A sufficiently suitable application of a conceptual model in the context of the planetary boundaries as in Steffen et al. (2015) requires the model's ability to simulate credible transients of global carbon dynamics. In order to achieve this, we calibrated model parameters to observed carbon fluxes and pools, as well as simulation results of detailed high-resolution Earth system models.

Because we simulate relative dynamics between the different carbon compartments and do not aim at prognostics of actual time evolution of carbon pools, all carbon fluxes and pools are normalised to the active carbon at pre-industrial times, i.e. the total sum of pre-industrial carbon in the year 1750 AD (3989 Gt C, Fig. 1). All normalised parameter values are summarised in Table 2.

2.2.1 Temperature

For the calibration of the linear relationship between temperature and atmospheric carbon content (Eq. 10) we used the transient climate response to cumulative emissions (TCRE) with a reported global mean surface temperature increase per emitted carbon of 2 K / 1000 Gt C (Joos et al., 2013; Gillett et al., 2013). Assuming an airborne fraction of 0.5 (Knorr, 2009; Gloor et al., 2010), the global mean temperature increase rate per atmospheric carbon increase (Eq. 10) is approximately twice the temperature increase rate of emitted carbon (TCRE), i.e. 2 K / 500 Gt C in the atmosphere. From this global surface temperature increase rate (two-thirds ocean and one-third land surface), the global land surface temperature increase can be inferred via the global land / sea warming ratio of approximately 1.6 (Sutton et al., 2007). Thus, we approximate a global land surface warming rate of 5.3 K / 1000 Gt C that remains in the atmosphere. The y-offset (b_T in Eq. 10) was inferred via global land surface temperature anomalies from 1880–2000 (Jones et al., 2012), a global average (1880–2000) land temperature of 8.5 °C (NOAA, 2015) and observed monthly mean CO_2 concentrations (Mauna Loa, 1959–2000, Tans and Keeling, 2015).

2.2.2 Ocean–atmosphere dynamics

The carbon solubility in sea water factor (β) is directly determined by the assumption of pre-industrial equilibrium between upper-ocean carbon and atmospheric carbon ($\dot{C_\mathrm{m}}(0) = 0$). From this and a present carbon flux from the atmosphere to the ocean of $\dot{C_\mathrm{m}}(t_\mathrm{tod}) = 2.3\,\mathrm{Gt\,C\,a^{-1}}$ (Ciais et al., 2013), follows the atmosphere–ocean diffusion coefficient a_m.

2.2.3 Terrestrial dynamics

Photosynthesis and respiration are calibrated according to temperature relationships reported in the literature. However, literature generally specifies temperature relationships at small temporal- and spatial-scales in controlled environments, whereas our model equations refer to a global average of day and night-time temperature. Thus, only a rough estimation of the relationship between temperature and photosynthesis / respiration for model calibration is possible. As in Anderies et al. (2013), we assume maximum respiration at a global land surface temperature of 18 °C (supported by Yuan et al. (2011)), determining the ratio of parameters $b_\mathrm{r}/c_\mathrm{r} = 18\,°\mathrm{C}$ (Fig. 3). We choose a maximum of photosynthesis at 12 °C, incorporating a CO_2 fertilisation feedback indirectly via the dependence of temperature on atmospheric carbon ($b_\mathrm{p}/c_\mathrm{p} = 12\,°\mathrm{C}$). The amplitudes of photosynthesis and respiration functions (a_r and a_p, respectively) are approximated for agreement with carbon fluxes reported in Ciais et al. (2013). Note that the functional form of carbon fluxes is not decisive for the model dynamics, however, it is important that the curves of photosynthesis and respiration intersect at some temperature limit where ecosystem respiration exceeds photosynthesis. With our parameterisation this is the case at a global mean land surface temperature of approximately 13 °C, which is 4.5 °C warmer than the 20th century average global mean land surface temperature (NOAA, 2015). This is in line with multi-model assessments in carbon reversal studies (e.g. Heimann and Reichstein, 2008; Friend et al., 2013).

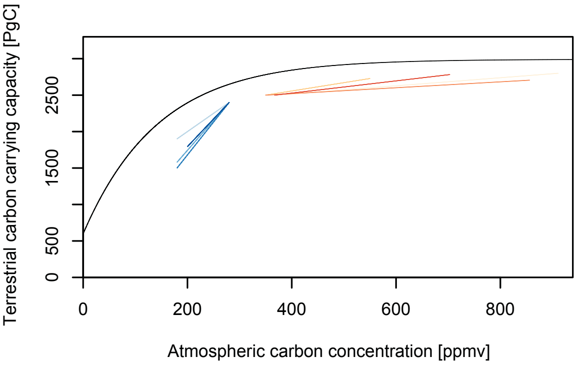


Figure 4. Approximated terrestrial carbon carrying capacity (black line). Blue lines represent approximate changes in terrestrial carbon storage published in Crowley (1995), François et al. (1998), Kaplan et al. (2002) and Joos et al. (2004). Red lines represent simulated changes in terrestrial carbon storage due to climate change reported by Joos et al. (2001), Lucht et al. (2006) and Friend et al. (2013).

The terrestrial carbon carrying capacity $K(C_\mathrm{a}(t))$ in $\dot{C}_t(t)$ determines how much carbon can be accumulated in the terrestrial system at maximum, as long as photosynthesis exceeds respiration (refer to Eq. 6). $K(C_\mathrm{a}(t))$ was calibrated to represent both past long-term climatic and terrestrial carbon changes (last glacial maximum to Holocene) (Crowley, 1995; François et al., 1998; Kaplan et al., 2002; Joos et al., 2004), and prognostics of climate change impacts on terrestrial carbon storage (Joos et al., 2001; Lucht et al., 2006; Friend et al., 2013) to capture terrestrial changes due to climate variability (Fig. 4).

Human activities such as fires, deforestation and agricultural land use that affect terrestrial carbon stocks are summarised as human offtake of biomass and are presently estimated at $H(t_\mathrm{tod}) = 1.1\,\mathrm{Gt\,C\,a^{-1}}$ (Ciais et al., 2013). With a present terrestrial carbon pool of $C_\mathrm{t}(t_\mathrm{tod}) = 2470\,\mathrm{Gt\,C}$ we calculate the human offtake rate $\alpha = H(t_\mathrm{tod})/C_\mathrm{t}(t_\mathrm{tod})$.

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

Table 2. Calibrated model parameters after normalisation to pre-industrial carbon pools. Remaining units are years (a) and temperature (20 K).

Parameter	Symbol	Value	Unit
Ecosystem-dependent conversion factor	r_{tc}	2.5	a^{-1}
Scaling factor for photosynthesis $P(T)$	a_{p}	0.48	$(20\,\text{K})^{-b_{\text{p}}}$
Scaling factor for respiration $R(T)$	a_{r}	0.40	$(20\,\text{K})^{-b_{\text{r}}}$
Power law exponent for increase in $P(T)$ for low T	b_{p}	0.5	1
Power law exponent for increase in $R(T)$ for low T	b_{r}	0.5	1
Rate of exponential decrease in $P(T)$ for high T	c_{p}	0.556	$(20\,\text{K})^{-1}$
Rate of exponential decrease in $R(T)$ for high T	c_{r}	0.833	$(20\,\text{K})^{-1}$
Scaling factor for terrestrial carbon carrying capacity	a_k	−0.6	1
Rate of exponential increase in terrestrial carbon carrying capacity	b_k	13.0	1
Offset for terrestrial carbon carrying capacity	c_k	0.75	1
Human terrestrial carbon offtake rate	α	0.0004	a^{-1}
Slope of T–C_{a} relationship	a_T	1.06	20 K
Intercept of T–C_{a} relationship	b_T	0.227	20 K
Carbon solubility in sea water factor	β	0.654	1
Atmosphere–ocean diffusion coefficient	a_m	0.0166	20 K
* Atmospheric carbon threshold of tCDR implementation	$\widetilde{C_{\text{a}}}$	0–0.3	1
Rapidity of tCDR ramp-up (tCDR implementation capacity)	s_{CE}	200	1
* Maximum tCDR rate	α_{max}	0–0.03	a^{-1}
* Size of geological fossil carbon stock	c_{max}	0–0.51	1
Industrialisation rate	r_i	0.03	a^{-1}
Climate change boundary	b_{a}	0.21	1
Land system change boundary	b_{l}	0.59	1
Ocean acidification boundary	b_{m}	0.31	1

* Parameters are varied during the analysis and the parameter range is stated.

2.2.4 Fossil fuel emissions

The size of the geological fossil carbon stock c_{max} determines the carbon released from fossil reservoirs (Eq. 2) and plays an important role for carbon dynamics (Sect. 3.4). In the scope of this study, c_{max} is varied to assess different baseline emissions following the cumulative emissions of the representative concentration pathways (RCPs). RCP2.6 is a low-emission scenario with cumulative emissions of approximately 880 Gt C ($c_{\text{max}} = 0.2$) (van Vuuren et al., 2011). The two medium emission scenarios RCP4.5 and RCP6.0 have cumulative emissions of approximately 1200 Gt C ($c_{\text{max}} = 0.31$) (Thomson et al., 2011) and 1400 Gt C ($c_{\text{max}} = 0.36$) (Masui et al., 2011), respectively. RCP8.5 represents a business as usual scenario with cumulative emissions of approximately 2000 Gt C ($c_{\text{max}} = 0.51$) (Riahi et al., 2011).

2.3 Planetary boundaries

We use the carbon-related planetary boundaries (climate change, ocean acidification and land system change) to define the desirability of given trajectories of carbon pool evolution. The proposed locations of these boundaries are normalised to match the normalisation of our model.

The planetary boundary for climate change is proposed at 350 ppmv CO_2 equivalents in the atmosphere with an uncertainty range to 450 ppmv (Steffen et al., 2015). For our study we take the middle of the uncertainty range (400 ppmv) because critical atmospheric thresholds are likely to be located somewhere within the uncertainty range and obtain a normalised climate change boundary is at 0.21 atmospheric carbon. Ocean acidification is measured via the saturation state of aragonite and its boundary is set at 80 % of the pre-industrial average annual global saturation state of aragonite (Steffen et al., 2015). Since chemical processes are not explicitly represented in our model, this measure is not directly transferable to maritime carbon content. This measure is not directly transferable to maritime carbon content because it largely depends on chemical variables such as pH-value, ocean alkalinity and dissolved inorganic carbon that are not included in the model. At the current carbon content (1150 Gt C), the saturation state of aragonite is at 84 % of the pre-industrial value (Guinotte and Fabry, 2008). We therefore estimate the normalised ocean acidification boundary at 0.31, about 5 % higher than the current value of the marine

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

carbon stock (0.29). The land system change boundary is defined in terms of the amount of remaining forest cover, motivated by critical biogeophysical feedbacks of forest biomes to the physical climate system (Steffen et al., 2015). The global boundary has been specified as 75 % of global forest cover remaining (Steffen et al., 2015). Due to the lack of biogeophysical feedbacks in the model, we translate deforestation into carbon content by measuring the loss of vegetation carbon with deforestation. We thereby neglect vegetation carbon of all non-forest biomes, while at the same time neglecting soil carbon changes by deforestation (Heck et al., 2016), thus approximating that soil carbon changes by deforestation are of the same order of magnitude as vegetation carbon pools of non-forest biomes. With vegetation carbon of 550 Gt C (Ciais et al., 2013), we obtain a normalised land system change boundary at 0.59.

Note that the exact location and normalisation of the boundaries is not decisive for our results because we qualitatively analyse the influence of tCDR management on the existence of desirable trajectories. Slightly different sets of planetary boundaries would not qualitatively change the systemic effects reported in this study.

2.4 Model analysis and terminology

Our analysis of the co-evolutionary system aims at assessing transient dynamics of carbon pools with respect to planetary boundaries. First, we run the model and exemplarily show the influence of socially controlled parameters of tCDR implementation on the transient carbon pool evolution (Sect. 3.1). It is of particular relevance under what circumstances the simulated carbon pool trajectories (atmosphere, ocean and land) do not cross their respective planetary boundaries. We refer to the regions on the safe side of the planetary boundaries as “safe regions”. All carbon pool trajectories remaining in the respective safe region at all times are considered “safe trajectories”. For example, all atmospheric carbon trajectories that do not cross the planetary boundary for climate change (i.e. trajectories that are in the safe region of atmospheric carbon) are safe atmospheric carbon trajectories. System states with each carbon pool remaining in its respective safe region are referred to as carbon system states within the SOS, i.e. “safe states”.

In a nonlinear dynamical system, trajectories can be sensitive to initial conditions. The pre-industrial distribution of carbon pools, as well as carbon dynamics in the Earth system are relatively well-assessed, while still subject to high uncertainty (Ciais et al., 2013). Furthermore, considerable uncertainty remains with respect to our conceptual model structure and the exact values of planetary boundaries. Bearing in mind these inherent uncertainties, we explore how robust the existence of safe trajectories is under a variation of the initial conditions, i.e. the initial carbon pool distribution and different tCDR characteristics (Sect. 3.2).

Such a variation of initial conditions is also a common approach to conceptualising and measuring resilience of social–ecological systems as the ability to return to an attracting state after a perturbation (Holling, 1973; Scheffer et al., 2001). A suitable approach to quantifying the likelihood of a complex system to return to an attracting state under finite perturbations is basin stability analysis (Menck et al., 2013).

In the context of planetary boundaries, not necessarily all trajectories that approach a “safe attractor” (i.e. an attractor within the SOS associated to all three planetary boundaries) would be considered safe because they could temporarily leave the safe region. The concept of constrained basin stability (van Kan et al., 2016) and related methods (Hellmann et al., 2016) provide generalisations of basin stability that allow taking transient phenomena into account. Similarly to the constrained basin stability approach, we classify different domains in the initial-condition state space based on transient dynamics of carbon pools. The set of initial conditions resulting in safe carbon trajectories form the “safe domain”. We refer to this domain as the manageable core of the safe operating space (MCSOS), as it depends on the tCDR management characteristics and the emission pathway. The “undesirable domain” is formed by all initial conditions resulting in a transgression of all three carbon boundaries at some point in time. Remaining state space domains are formed by initial conditions leading to a transgression of a subset of planetary boundaries. They are referred to as the respective partially manageable domains (MDs) (e.g. the land manageable domain is the state space domain of initial conditions with trajectories without a transgression of the land boundary).

The computational efficiency of our model allows for a systematic analysis of the MCSOS and other domains under variation of societal parameters (tCDR management and fossil fuel emissions). We analyse how the size of all domains (MCSOS, partially MDs and the undesirable domain) varies with different tCDR characteristics (Sect. 3.3) and emission pathways (Sect. 3.4). In the spirit of van Kan et al. (2016), the size of (partially) manageable domains can be interpreted as a resilience-like measure of the opportunities to stay within the carbon-related SOS, taking into account inherent structural uncertainties of our model, the location of planetary boundaries, and the pre-industrial carbon pool distribution. Note that the maximum extent of the MCSOS is constrained by the planetary boundaries, but it may differ from the SOS (i.e. the “safe” region) as the safety of the domain is determined by transient system dynamics, whereas the SOS is defined within static planetary boundaries.

3 Results and Discussion

3.1 Carbon system trajectories subject to societal tCDR management loop

To illustrate how the co-evolutionary social–environmental system evolves with respect to carbon-related planetary

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

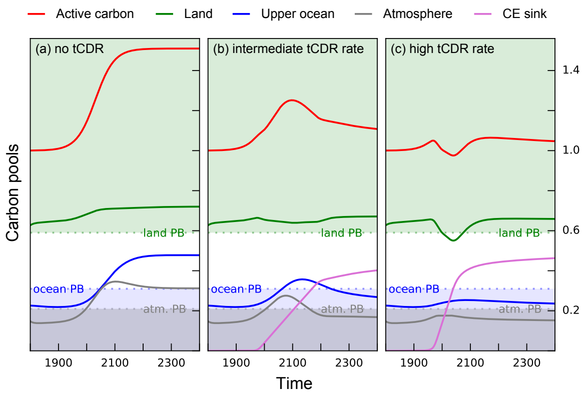


Figure 5. Time evolution of the normalised carbon pools in our model of the carbon system for three tCDR configurations with a high-emission baseline (cumulative emissions as in RCP8.5; Riahi et al., 2011) **(a)** without tCDR ($\alpha_\mathrm{max} = 0$), **(b)** intermediate tCDR rate ($\alpha_\mathrm{max} = 0.0025$) and **(c)** high tCDR rate ($\alpha_\mathrm{max} = 0.025$). Total active carbon (red) is increased by fossil fuel emissions ($c_\mathrm{max} = 0.51$) with dynamic response of the terrestrial carbon pool (green), maritime carbon pool (blue) and atmospheric carbon pool (grey). The tCDR sink (purple) stores carbon extracted from the active system. Shaded areas represent the respective safe regions of land, ocean and atmosphere in green, blue and grey. Dotted lines indicate the location of the associated planetary boundaries (PBs).

boundaries, Fig. 5 depicts trajectories of the major carbon pools with tCDR adhering to different management characteristics. All trajectories start at their respective normalised pre-industrial state. The normalised planetary boundaries (Sect. 2.3) are indicated as dotted lines and the safe region of each boundary (refer to Sect. 2.4) is shaded in the respective colours. Variation of tCDR characteristics reflects uncertainty about possible tCDR rates related to overall biomass harvesting potentials and societies' implementation capacities (Sect. 2.1).

The emission baseline used for all results displayed in Fig. 5 is a business-as-usual scenario with cumulative emissions as in RCP8.5 (Riahi et al., 2011). Without tCDR (Fig. 5a), all that fossil carbon societies emit into the atmosphere is distributed to ocean, land and atmosphere. This results in more active carbon (red), leading to carbon accumulation in all pools and a transgression of the atmosphere and ocean boundaries. In this emission scenario, the land system accumulates carbon and, thus, moves away from its planetary boundary in our model setting (note that the actual control variable of the planetary boundary of land system change as defined by Steffen et al. (2015) is the remaining forest cover, which would not be directly modified by changing atmospheric carbon concentrations). Moreover, higher emission baselines (results not shown here) can lead to decreasing terrestrial carbon stocks when respiration dominates over photosynthesis due to strong global warming.

In Fig. 5b) and c), the societal tCDR response via harvesting from the terrestrial carbon stock and subsequent storage starts just before the atmospheric boundary is reached ($\widetilde{C_\mathrm{a}} = 0.18 \sim 340$ ppmv). With a low tCDR rate (maximal storage flux of about 7 Gt C a^{-1}, $\alpha_\mathrm{max} = 0.0025$), the CE sink is filled relatively slowly (Fig. 5b). Thus, a transient transgression of the atmosphere and ocean boundaries cannot be prevented. However, all trajectories re-enter their respective "safe" region after about 150 years. A higher tCDR rate ($\alpha_\mathrm{max} = 0.025$, corresponding to very high-potential storage fluxes of 26 Gt C a^{-1} or 5 % of global biomass per year) can prevent a large increase in active carbon and thus prevents the transgression of both atmosphere and ocean boundaries (Fig. 5c). However, extensive harvest from the land carbon pool then leads to a temporary transgression of the land boundary. The implementation of tCDR was thus effective in its purpose of preventing entry into a dangerous region of climate change, but at the cost of exploiting the land system to an extent that crossed the land system change boundary.

These results show that small tCDR rates (Fig. 5b) (or implementation that is too late, results not shown here) do not necessarily keep the system in the SOS. High tCDR rates (Fig. 5c) could seem successful when focusing on the climate change boundary, but might in fact not be feasible if other components of the carbon system are taken into account. In light of ongoing deforestation for the purpose of bioenergy production (Gao et al., 2011), this simulated collateral transgression of the land system change boundary with large-scale tCDR is an important and plausible feature of the model.

In the actual Earth system, a transgression of the land system change boundary might evoke additional trade-offs to the biogeophysical climate system (Foley et al., 2003), which are not represented in the model. For example, large tCDR rates can only be achieved by large-scale land-use change that could alter atmospheric circulations and rainfall patterns (Snyder et al., 2004) even though the carbon-related climate change boundary might not be transgressed with high tCDR rates.

The carbon values stated here are primarily given as an orientation for the reader, and should not be directly interpreted with respect to tCDR feasibility assessments. However, tCDR rates of 7 Gt C a^{-1} are in line with more conservative biomass harvest potentials considering biodiversity conservation and agricultural limits (Dornburg et al., 2010; Beringer et al., 2011). More idealistic assessments of tCDR rates of more than 35 Gt C a^{-1} – assuming high biomass yields of more than one-quarter of global land area – have been reported as well (Smeets et al., 2007). In this context, the range of tCDR rates studied in this paper reflects both conservative and highly optimistic tCDR potentials reported in the literature.

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

3.2 State space domain structure of the Earth's carbon system subject to societal tCDR management loop

We compute the state space domain structure (refer to Sect. 2.4) from a sample of initial conditions around the pre-industrial carbon state. We sample approximately 66 000 initial conditions from a regular grid by variation of each carbon pool by ±0.2 around the pre-industrial conditions. This range is a pragmatic choice which does not influence the following qualitative analysis. To compute the existing domains, we evolve each initial condition for 600 years in time and colour it according to the domains following from the transient properties of the trajectories of land, atmosphere and ocean carbon, as described above. The mapping of initial conditions sheds light on possible domains in the carbon system and potential transitions into other state space domains in our model of the carbon cycle. In this context, the vicinity of the pre-industrial and current Earth system states to such domain boundaries in the model's initial-carbon-condition state space is of particular relevance.

Figure 6 shows the existing domains without tCDR (a), with intermediate tCDR rates (b) and with very high tCDR rates (c). The emission baseline is the same for all variations of tCDR characteristics, with cumulative emissions of approximately 880 Gt C, which is comparable to RCP2.6 cumulative emissions (van Vuuren et al., 2011). The current state of the carbon cycle is located in proximity to domain borders, highlighting that it is close to a transgression of the land system and climate change boundaries. Historical emissions and land system changes have moved the state of the carbon cycle closer towards the undesirable domain, and remaining on an emission trajectory similar to RCP2.6 without tCDR results in the non-existence of the MCSOS (Fig. 6a). Thus, the manageable core does not exist if the implementation of tCDR management is not considered by society, even in a relatively low-emission scenario.

Figure 6b and c serve as an example of how human intervention and management by tCDR can influence the size and even the existence of the MCSOS and other domains. With an implementation of tCDR, the MCSOS can be re-established, potentially to its full extent, which is directly determined by the three planetary boundaries (Fig. 6b). Even for a relatively low-emission scenario, the tCDR threshold needs to be at sufficiently low atmospheric carbon content ($\widetilde{C_a} = 0.16$) to prevent potential boundary transgressions. Nevertheless, because of past land-use change, the current Earth system state is approaching domains with unsafe land system and climate change. If tCDR is applied under the same conditions but with a 10 times higher potential tCDR rate ($\alpha_{\max} = 0.04$), the MCSOS shrinks due to over-exploitation of the land system for tCDR (Fig. 6c). The land system is overexploited when the total human biomass offtake flux ($H_{CE} + H$) exceeds net ecosystem productivity (NEP). This decreases terrestrial carbon pools (Eq. 5) which in turn limits the potential for tCDR (Eq. 11). In Fig. 6c this occurs under high initial atmospheric carbon concentrations, because these result in a higher tCDR flux for the same potential tCDR rate ($\alpha_{\max}$, ref. to Fig. 2). The current state of the carbon cycle of the Earth system is out of the MCSOS. In this case, large societal commitment to avoid a transgression of the climate change boundary leads to a collateral transgression of the land system change boundary in our model.

3.3 Size of manageable domains under variation of tCDR characteristics

The size and existence of the MCSOS and other state space domains depends on tCDR characteristics (refer to Sect. 2.4). We compute the size of the different initial-condition state space domains depending on the most decisive management parameters, i.e. on the implementation threshold $\widetilde{C_a}$ and on the potential maximum tCDR rate $\alpha_{\max}$. The size of all domains is measured in relation to the size of the considered state space section as depicted in Fig. 6, which is given by a variation of pre-industrial conditions by ±0.2.

Figure 7 depicts the relative size of the MCSOS and the partially manageable domains under baseline emissions of $c_{\max} = 0.4$, corresponding to cumulative emissions in the order of RCP6.0. The size of the MCSOS or partially MDs can be interpreted as a form of resilience of the system (i.e. the likelihood that the system stays within the carbon-related SOS). Thus, we measure the resilience of the carbon cycle by the size of MCSOS (i.e. the opportunity of success of tCDR to maintain safe trajectories). This strongly depends on the atmospheric carbon threshold at which tCDR is implemented. Obviously, only the anticipation of an approaching planetary boundary can prevent a transgression thereof. Thresholds higher than the atmospheric carbon boundary ($b_l = 0.21$) are not sufficient in sustaining a MCSOS, because the atmosphere MD disappears by definition at $\widetilde{C_a} = 0.21$ (grey line in Fig. 7a).

However, strong anticipation coupled with too early tCDR implementation does not necessarily maintain the system within the SOS. If tCDR is initialised at relatively low atmospheric carbon content ($\widetilde{C_a} = 0.13$ (approximately 330 ppmv) in Fig. 7a), the MCSOS is diminished due to a transgression of the land system change boundary at some point in time. Hence, the window of opportunity for using tCDR as a means of staying in the SOS under this exemplary fossil fuel emission scenario is limited to a relatively narrow range of tCDR implementation thresholds. The size of the land MD shows nonlinear dependence on the tCDR threshold. For thresholds between 0.2 and 0.25, the land MD is almost diminished (Fig. 7a), because the relatively high tCDR rate ($\alpha_{\max} = 0.02$) leads to an over-exploitation of the land system (ref. to Sect. 3.2). However, higher tCDR thresholds avoid this over-exploitation and increase the land MD, because of a later onset of tCDR and overall higher NEP due to higher atmospheric carbon content and temperature (Eq. 6).

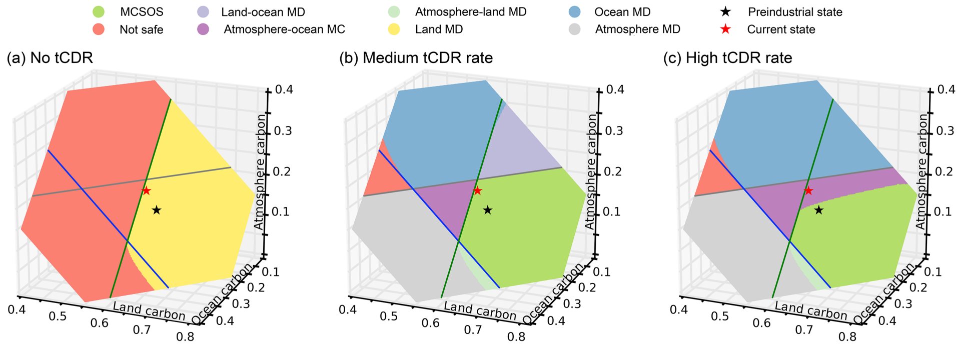


Figure 6. Charting of normalised carbon-system initial-condition state space in our model for three tCDR management characteristics with identical, relatively low-emission baseline ($c_{\max} = 0.2$): **(a)** without tCDR ($\alpha_{\max} = 0$), **(b)** intermediate tCDR rates ($\alpha_{\max} = 0.004$) and **(c)** high tCDR rates ($\alpha_{\max} = 0.04$). The two-dimensional plane is formed by sampling initial conditions around the pre-industrial state (variation of carbon stocks by ±0.2 while conserving total carbon in the system). Each domain is coloured according to transient properties of trajectories starting in different state space regions. For example, the MCSOS (i.e. safe domain) is formed by the initial conditions of "safe" trajectories, whereas red indicates the initial conditions of trajectories crossing all respective planetary boundaries at some point in the simulation. Lines indicate the associated planetary boundaries of atmosphere, land and ocean in grey, green and blue, respectively.

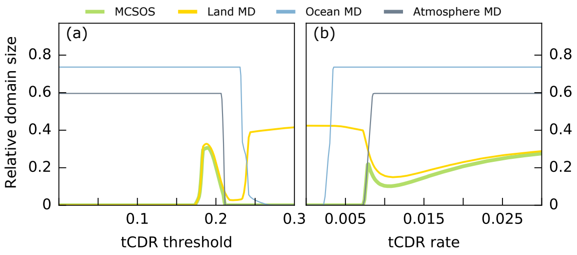


Figure 7. Relative size of domains in modelled carbon-system initial-condition state space for normalised parameter variation of **(a)** tCDR threshold (with $\alpha_{\max} = 0.02$) and **(b)** tCDR rate (with $\widetilde{C_a} = 0.2$) for a medium emission scenario ($c_{\max} = 0.4 \sim 1600\,\mathrm{Gt\,C}$ cumulative emissions). All domain sizes are given as shares of the state space region defined by a variation of the pre-industrial conditions by ±0.2.

Similar to the tCDR threshold, the parameter governing the maximal achievable rate of tCDR plays a decisive role for the existence of the MCSOS. With a tCDR implementation threshold not far below the atmospheric carbon boundary ($\widetilde{C_a} = 0.2$), high tCDR rates are required in order to maintain a MCSOS. The tCDR starts being effective in maintaining a MCSOS at a rate of $\alpha_{\max} > 0.007$ (corresponding to approximately $16.5\,\mathrm{Gt\,C\,a^{-1}}$ with a fixed land carbon pool of 0.6). Rates smaller than that are not sufficient because of a lacking atmospheric MD (grey line in Fig. 7b).

As the tCDR threshold, the tCDR rate has a strong influence on the size of the land MD. For small tCDR rates, the land MD is sustained because of high atmospheric carbon concentrations and small biomass extraction. Rates higher than $\alpha_{\max} = 0.0075$ result in a smaller land MD due to the over-exploitation of the photosynthetic productivity of the system which is reduced by both biomass removal and decreasing atmospheric carbon concentrations driving NEP. Higher rates, however, lead to overall smaller reductions of the land MD. This nonlinearity is evoked by the co-evolutionary feedbacks between society and the carbon cycle, which lead to a deceasing tCDR flux if the system is in the atmosphere MD. Thus, sufficiently high tCDR rates lead to fast atmospheric carbon decrease and tCDR is switched off before the land system boundary is transgressed.

This analysis of the size of initial-condition state space domains suggests that the success of tCDR in sustaining the Earth system's persistence in the carbon SOS nonlinearly depends on the characteristics of tCDR implementation. On the one hand, foresightedness and anticipation of planetary boundaries are required to maintain the MCSOS, while on the other hand, too-early or too-intensive management could trigger co-transgressions of other planetary boundaries.

3.4 Opportunities and limitations of tCDR

While anticipation and appropriate management are necessary, the underlying emission scenario plays a major role in the resulting carbon dynamics. Figure 8 exemplarily depicts the relative MCSOS size for variations of tCDR characteristics (threshold and potential maximum rate) for emission pathways in accordance with RCP cumulative-emission scenarios. The window of opportunity for successful tCDR (i.e. the size of the MCSOS) decreases with increasing emission baselines and depends on the tCDR rate and threshold. In the case of the low-emission RCP2.6 scenario ($c_{\max} = 0.2$), the MCSOS can be sustained for a broad range of parameter values (Fig. 8a). The medium emission scenarios RCP4.5 ($c_{\max} = 0.31$; Thomson et al., 2011) and RCP6.0 ($c_{\max} =$

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

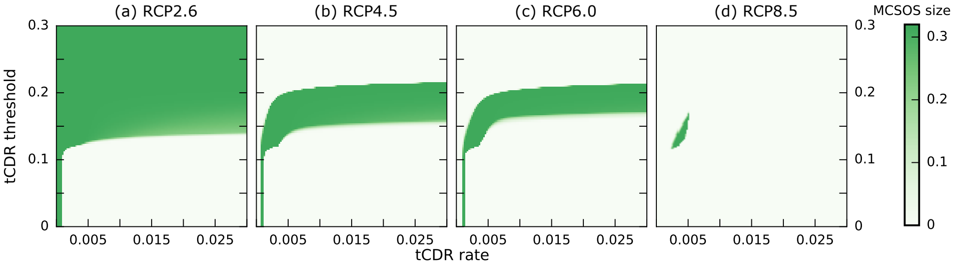


Figure 8. Relative size of the MCSOS for normalised parameter variation of potential maximum tCDR rate (x axis) and tCDR threshold (y axis) for different underlying emission scenarios: **(a)** RCP2.6 ($c_{\max} = 0.2$), **(b)** RCP4.5 ($c_{\max} = 0.31$), **(c)** RCP6.0 ($c_{\max} = 0.36$) and **(d)** RCP8.5 ($c_{\max} = 0.51$).

0.36; Masui et al., 2011) show a narrower range of tCDR characteristics that have the potential to sustain a MCSOS (Fig. 8b and c). In a business-as-usual RCP8.5 scenario, the room for manoeuvring to maintain a MCSOS is very small (Fig. 8d).

Besides the dependence on the emission scenario, Fig. 8 highlights that for most emission scenarios the range of tCDR thresholds sustaining the MCSOS is narrow and depends on the tCDR rate. As discussed in Sect. 3.3 (for a fixed tCDR rate), tCDR thresholds higher than the atmospheric carbon boundary (0.21) are not sufficient in preventing a boundary transgression in the medium-to-high emission scenarios (Fig. 8b–d), whereas small tCDR thresholds lead to a transgression of the land system change boundary (unless tCDR rates are within a very narrow range smaller than 0.001). The variation of both the tCDR rate and threshold shows that smaller tCDR rates require a smaller minimal tCDR threshold as well as a smaller maximal threshold (Fig. 8b–d). This dependence of the success of tCDR on both the tCDR characteristics and the underlying emission scenarios highlights the relevance of societal intervention for global carbon dynamics. Essentially, tCDR intervention can trigger a nonlinear carbon system response through the land system when human carbon offtake exceeds NEP, which in turn causes a further reduction in NEP and tCDR potentials.

In our conceptual framework, tCDR can be effective in complementing climate change mitigation strategies as employed in low-emission scenarios. However, already an RCP4.5 emission scenario narrows the range of potentially successful management options significantly in comparison to RCP2.6 emissions. Under a business-as-usual pathway, tCDR cannot be applied to maintain a MCSOS in a resilient way. In contrast to prevailing reasoning of CE as an emergency action in case of dangerous climate change (Caldeira and Keith, 2010), tCDR would most likely not function as an emergency option under high-emission scenarios when additional sustainability dimensions reflected by other planetary boundaries are taken into account.

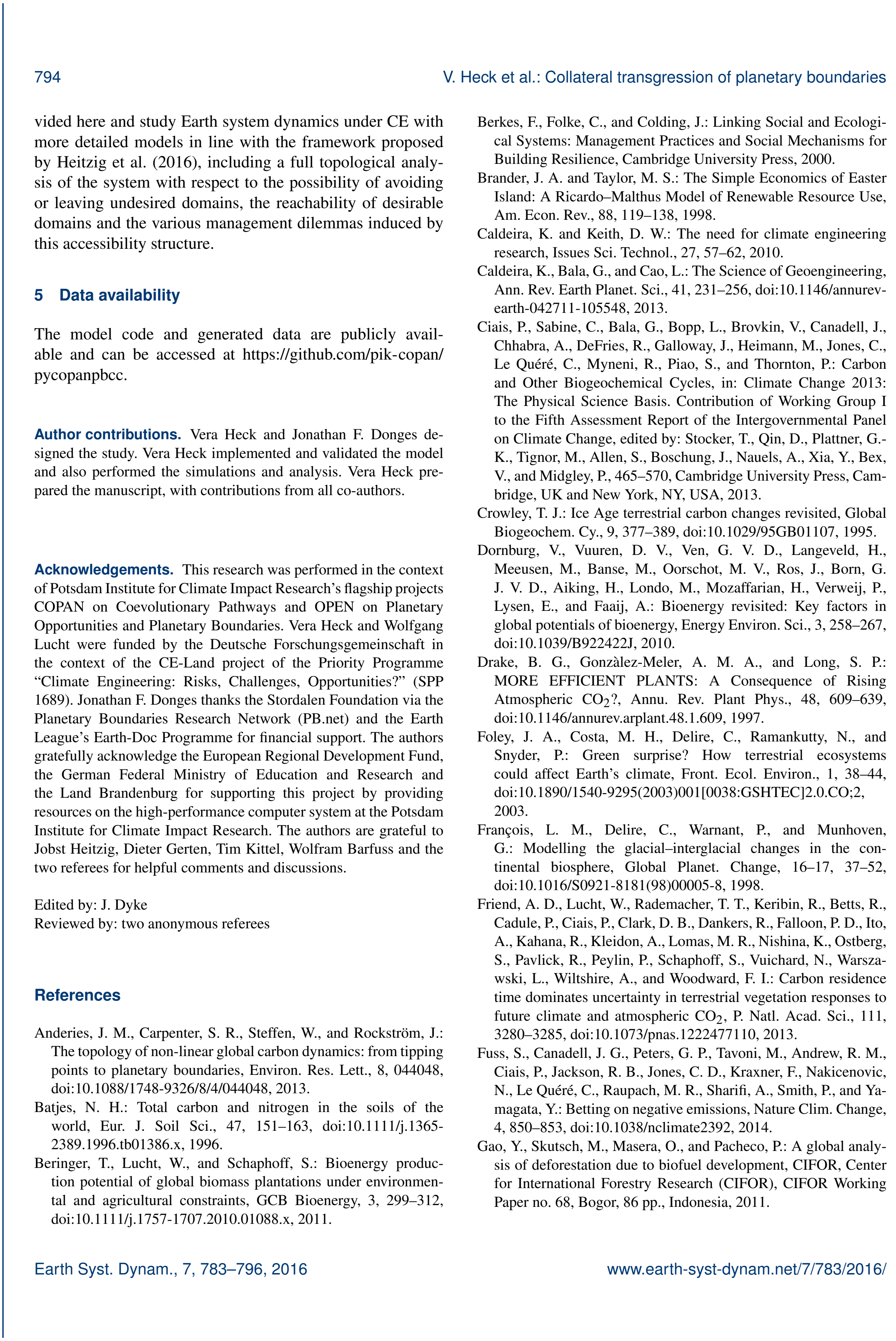
4 Conclusions

The introduced conceptual modelling approach – combining carbon cycle dynamics with a societal feedback loop of carbon monitoring and terrestrial carbon dioxide removal (tCDR) action – provides valuable insights into system-level constraints to navigating within the carbon-related safe operating space defined by several interlinked planetary boundaries. Despite the fact that the reported results cannot be taken as exact quantitative prognostics of carbon pool evolution, our analysis has shown that employing tCDR for managing the atmospheric carbon pool does not necessarily safeguard the carbon cycle in the safe operating space because of nonlinear feedbacks between tCDR management and the carbon system.

The success of maintaining a manageable core of the safe operating space depends on the degree of anticipation of climate change, the potential maximum tCDR rate, as well as the underlying emission pathway. While tCDR might be successfully deployed as part of a strong climate change mitigation scenario, it is not likely to be effective in a business-as-usual scenario. Particularly, the focus on one planetary boundary alone (e.g. climate change), may lead to navigating the Earth system out of the carbon-related safe operating space due to collateral transgression of other boundaries (e.g. land system change). In light of numerous (economically-based) integrated assessment studies proposing tCDR to counteract anthropogenic emissions, our conceptual results highlight that it is vital to include integrated sustainability assessments of more advanced models to the debate on climate engineering (CE) and climate change mitigation via tCDR. In the case of tCDR, the consequences for biosphere integrity, as well as trade-offs with agricultural land use and the biogeophysical climate system must be taken into account among other sustainability dimensions reflected by planetary boundaries and beyond.

In analogy to our analysis of tCDR, the approach followed in this paper could be transferred to other CE proposals such as ocean fertilisation or solar radiation management. Additionally, it would be of interest to extend the analysis pro-

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).



vided here and study Earth system dynamics under CE with more detailed models in line with the framework proposed by Heitzig et al. (2016), including a full topological analysis of the system with respect to the possibility of avoiding or leaving undesired domains, the reachability of desirable domains and the various management dilemmas induced by this accessibility structure.

5 Data availability

The model code and generated data are publicly available and can be accessed at https://github.com/pik-copan/pycopanpbcc.

Author contributions. Vera Heck and Jonathan F. Donges designed the study. Vera Heck implemented and validated the model and also performed the simulations and analysis. Vera Heck prepared the manuscript, with contributions from all co-authors.

Acknowledgements. This research was performed in the context of Potsdam Institute for Climate Impact Research's flagship projects COPAN on Coevolutionary Pathways and OPEN on Planetary Opportunities and Planetary Boundaries. Vera Heck and Wolfgang Lucht were funded by the Deutsche Forschungsgemeinschaft in the context of the CE-Land project of the Priority Programme "Climate Engineering: Risks, Challenges, Opportunities?" (SPP 1689). Jonathan F. Donges thanks the Stordalen Foundation via the Planetary Boundaries Research Network (PB.net) and the Earth League's Earth-Doc Programme for financial support. The authors gratefully acknowledge the European Regional Development Fund, the German Federal Ministry of Education and Research and the Land Brandenburg for supporting this project by providing resources on the high-performance computer system at the Potsdam Institute for Climate Impact Research. The authors are grateful to Jobst Heitzig, Dieter Gerten, Tim Kittel, Wolfram Barfuss and the two referees for helpful comments and discussions.

Edited by: J. Dyke
Reviewed by: two anonymous referees

References

Anderies, J. M., Carpenter, S. R., Steffen, W., and Rockström, J.: The topology of non-linear global carbon dynamics: from tipping points to planetary boundaries, Environ. Res. Lett., 8, 044048, doi:10.1088/1748-9326/8/4/044048, 2013.

Batjes, N. H.: Total carbon and nitrogen in the soils of the world, Eur. J. Soil Sci., 47, 151–163, doi:10.1111/j.1365-2389.1996.tb01386.x, 1996.

Beringer, T., Lucht, W., and Schaphoff, S.: Bioenergy production potential of global biomass plantations under environmental and agricultural constraints, GCB Bioenergy, 3, 299–312, doi:10.1111/j.1757-1707.2010.01088.x, 2011.

Berkes, F., Folke, C., and Colding, J.: Linking Social and Ecological Systems: Management Practices and Social Mechanisms for Building Resilience, Cambridge University Press, 2000.

Brander, J. A. and Taylor, M. S.: The Simple Economics of Easter Island: A Ricardo–Malthus Model of Renewable Resource Use, Am. Econ. Rev., 88, 119–138, 1998.

Caldeira, K. and Keith, D. W.: The need for climate engineering research, Issues Sci. Technol., 27, 57–62, 2010.

Caldeira, K., Bala, G., and Cao, L.: The Science of Geoengineering, Ann. Rev. Earth Planet. Sci., 41, 231–256, doi:10.1146/annurev-earth-042711-105548, 2013.

Ciais, P., Sabine, C., Bala, G., Bopp, L., Brovkin, V., Canadell, J., Chhabra, A., DeFries, R., Galloway, J., Heimann, M., Jones, C., Le Quéré, C., Myneni, R., Piao, S., and Thornton, P.: Carbon and Other Biogeochemical Cycles, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., 465–570, Cambridge University Press, Cambridge, UK and New York, NY, USA, 2013.

Crowley, T. J.: Ice Age terrestrial carbon changes revisited, Global Biogeochem. Cy., 9, 377–389, doi:10.1029/95GB01107, 1995.

Dornburg, V., Vuuren, D. V., Ven, G. V. D., Langeveld, H., Meeusen, M., Banse, M., Oorschot, M. V., Ros, J., Born, G. J. V. D., Aiking, H., Londo, M., Mozaffarian, H., Verweij, P., Lysen, E., and Faaij, A.: Bioenergy revisited: Key factors in global potentials of bioenergy, Energy Environ. Sci., 3, 258–267, doi:10.1039/B922422J, 2010.

Drake, B. G., Gonzàlez-Meler, A. M. A., and Long, S. P.: MORE EFFICIENT PLANTS: A Consequence of Rising Atmospheric CO_2?, Annu. Rev. Plant Phys., 48, 609–639, doi:10.1146/annurev.arplant.48.1.609, 1997.

Foley, J. A., Costa, M. H., Delire, C., Ramankutty, N., and Snyder, P.: Green surprise? How terrestrial ecosystems could affect Earth's climate, Front. Ecol. Environ., 1, 38–44, doi:10.1890/1540-9295(2003)001[0038:GSHTEC]2.0.CO;2, 2003.

François, L. M., Delire, C., Warnant, P., and Munhoven, G.: Modelling the glacial–interglacial changes in the continental biosphere, Global Planet. Change, 16–17, 37–52, doi:10.1016/S0921-8181(98)00005-8, 1998.

Friend, A. D., Lucht, W., Rademacher, T. T., Keribin, R., Betts, R., Cadule, P., Ciais, P., Clark, D. B., Dankers, R., Falloon, P. D., Ito, A., Kahana, R., Kleidon, A., Lomas, M. R., Nishina, K., Ostberg, S., Pavlick, R., Peylin, P., Schaphoff, S., Vuichard, N., Warszawski, L., Wiltshire, A., and Woodward, F. I.: Carbon residence time dominates uncertainty in terrestrial vegetation responses to future climate and atmospheric CO_2, P. Natl. Acad. Sci., 111, 3280–3285, doi:10.1073/pnas.1222477110, 2013.

Fuss, S., Canadell, J. G., Peters, G. P., Tavoni, M., Andrew, R. M., Ciais, P., Jackson, R. B., Jones, C. D., Kraxner, F., Nakicenovic, N., Le Quéré, C., Raupach, M. R., Sharifi, A., Smith, P., and Yamagata, Y.: Betting on negative emissions, Nature Clim. Change, 4, 850–853, doi:10.1038/nclimate2392, 2014.

Gao, Y., Skutsch, M., Masera, O., and Pacheco, P.: A global analysis of deforestation due to biofuel development, CIFOR, Center for International Forestry Research (CIFOR), CIFOR Working Paper no. 68, Bogor, 86 pp., Indonesia, 2011.

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

Gillett, N. P., Arora, V. K., Matthews, D., and Allen, M. R.: Constraining the Ratio of Global Warming to Cumulative CO_2 Emissions Using CMIP5 Simulations, J. Climate, 26, 6844–6858, doi:10.1175/JCLI-D-12-00476.1, 2013.

Gloor, M., Sarmiento, J. L., and Gruber, N.: What can be learned about carbon cycle climate feedbacks from the CO_2 airborne fraction?, Atmos. Chem. Phys., 10, 7739–7751, doi:10.5194/acp-10-7739-2010, 2010.

Guinotte, J. M. and Fabry, V. J.: Ocean Acidification and Its Potential Effects on Marine Ecosystems, Ann. NY Acad. Sci., 1134, 320–342, doi:10.1196/annals.1439.013, 2008.

Heck, V., Gerten, D., Lucht, W., and Boysen, L. R.: Is extensive terrestrial carbon dioxide removal a 'green' form of geoengineering? A global modelling study, Global Planet. Change, 137, 123–130, doi:10.1016/j.gloplacha.2015.12.008, 2016.

Heimann, M. and Reichstein, M.: Terrestrial ecosystem carbon dynamics and climate feedbacks, Nature, 451, 289–292, doi:10.1038/nature06591, 2008.

Heitzig, J., Kittel, T., Donges, J. F., and Molkenthin, N.: Topology of sustainable management of dynamical systems with desirable states: from defining planetary boundaries to safe operating spaces in the Earth system, Earth Syst. Dynam., 7, 21–50, doi:10.5194/esd-7-21-2016, 2016.

Hellmann, F., Schultz, P., Grabow, C., Heitzig, J., and Kurths, J.: Survivability of Deterministic Dynamical Systems, Scientific Reports, 6, 29654, doi:10.1038/srep29654, 2016.

Holling, C. S.: Resilience and Stability of Ecological Systems, Annu. Rev. Ecol. Syst., 4, 1–23, doi:10.1146/annurev.es.04.110173.000245, 1973.

Jarvis, A. J., Leedal, D. T., and Hewitt, C. N.: Climate-society feedbacks and the avoidance of dangerous climate change, Nature Clim. Change, 2, 668–671, doi:10.1038/nclimate1586, 2012.

Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., and Morice, C. P.: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010, J. Geophys. Res.-Atmos., 117, D05127, doi:10.1029/2011JD017139, 2012.

Joos, F., Prentice, I. C., Sitch, S., Meyer, R., Hooss, G., Plattner, G.-K., Gerber, S., and Hasselmann, K.: Global warming feedbacks on terrestrial carbon uptake under the Intergovernmental Panel on Climate Change (IPCC) Emission Scenarios, Global Biogeochem. Cy., 15, 891–907, doi:10.1029/2000GB001375, 2001.

Joos, F., Gerber, S., Prentice, I. C., Otto-Bliesner, B. L., and Valdes, P. J.: Transient simulations of Holocene atmospheric carbon dioxide and terrestrial carbon since the Last Glacial Maximum, Global Biogeochem. Cy., 18, GB2002, doi:10.1029/2003gb002156, 2004.

Joos, F., Roth, R., Fuglestvedt, J. S., Peters, G. P., Enting, I. G., von Bloh, W., Brovkin, V., Burke, E. J., Eby, M., Edwards, N. R., Friedrich, T., Frölicher, T. L., Halloran, P. R., Holden, P. B., Jones, C., Kleinen, T., Mackenzie, F. T., Matsumoto, K., Meinshausen, M., Plattner, G.-K., Reisinger, A., Segschneider, J., Shaffer, G., Steinacher, M., Strassmann, K., Tanaka, K., Timmermann, A., and Weaver, A. J.: Carbon dioxide and climate impulse response functions for the computation of greenhouse gas metrics: a multi-model analysis, Atmos. Chem. Phys., 13, 2793–2825, doi:10.5194/acp-13-2793-2013, 2013.

Kaplan, J. O., Prentice, I. C., Knorr, W., and Valdes, P. J.: Modeling the dynamics of terrestrial carbon storage since the Last Glacial Maximum, Geophys. Res. Lett., 29, 31-1–31-4, doi:10.1029/2002GL015230, 2002.

Keenan, T. F., Hollinger, D. Y., Bohrer, G., Dragoni, D., Munger, J. W., Schmid, H. P., and Richardson, A. D.: Increase in forest water-use efficiency as atmospheric carbon dioxide concentrations rise, Nature, 499, 324–327, doi:10.1038/nature12291, 2013.

Kellie-Smith, O. and Cox, P. M.: Emergent dynamics of the climate-economy system in the Anthropocene, Phil. Trans. A, 369, 868–86, doi:10.1098/rsta.2010.0305, 2011.

Kirtman, B., Power, S., Adedoyin, J., Boer, G., Bojariu, R., Camilloni, I., Doblas-Reyes, F., Fiore, A., Kimoto, M., Meehl, G., Prather, M., Sarr, A., Schär, C., Sutton, R., van Oldenborgh, G., Vecchi, G., and Wang, H.: Near-term Climate Change: Projections and Predictability, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., 953–1028, Cambridge University Press, Cambridge, UK and New York, NY, USA, 2013.

Knorr, W.: Is the airborne fraction of anthropogenic CO_2 emissions increasing?, Geophys. Res. Lett., 36, L21710, doi:10.1029/2009GL040613, 2009.

Kriegler, E., Hall, J. W., Held, H., Dawson, R., and Schellnhuber, H. J.: Imprecise probability assessment of tipping points in the climate system, P. Natl. Acad. Sci., 106, 5041–5046, doi:10.1073/pnas.0809117106, 2009.

Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J.: Tipping elements in the Earth's climate system, P. Natl. Acad. Sci., 105, 1786–1793, doi:10.1073/pnas.0705414105, 2008.

Lucht, W., Schaphoff, S., Erbrecht, T., Heyder, U., and Cramer, W.: Terrestrial vegetation redistribution and carbon balance under climate change, Carbon Balance and Management, 1, 1–6, doi:10.1186/1750-0680-1-6, 2006.

MacMartin, D. G., Kravitz, B., Keith, D. W., and Jarvis, A.: Dynamics of the coupled human–climate system resulting from closed-loop control of solar geoengineering, Clim. Dynam., 43, 243–258, doi:10.1007/s00382-013-1822-9, 2013.

Masui, T., Matsumoto, K., Hijioka, Y., Kinoshita, T., Nozawa, T., Ishiwatari, S., Kato, E., Shukla, P. R., Yamagata, Y., and Kainuma, M.: An emission pathway for stabilization at $6\,\mathrm{W\,m^{-2}}$ radiative forcing, Clim. Change, 109, 59–76, doi:10.1007/s10584-011-0150-5, 2011.

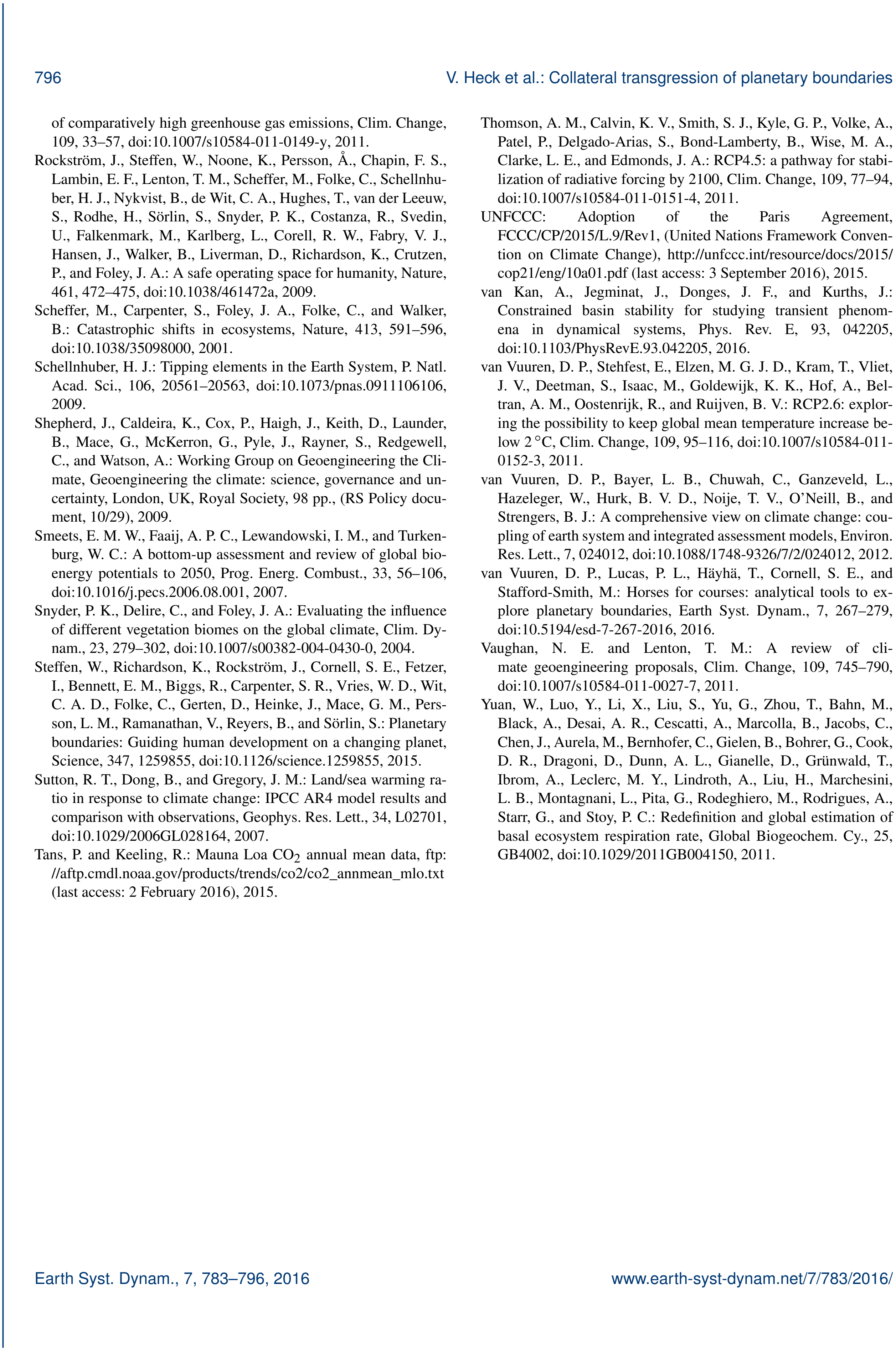
Menck, P. J., Heitzig, J., Marwan, N., and Kurths, J.: How basin stability complements the linear-stability paradigm, Nat. Phys., 9, 89–92, doi:10.1038/nphys2516, 2013.

Motesharrei, S., Rivas, J., and Kalnay, E.: Human and nature dynamics (HANDY): Modeling inequality and use of resources in the collapse or sustainability of societies, Ecol. Econ., 101, 90–102, doi:10.1016/j.ecolecon.2014.02.014, 2014.

NOAA National Centers for Environmental Information, State of the Climate: Global Analysis for Annual 2014, published online January 2015, available at: http://www.ncdc.noaa.gov/sotc/global/201413 (last access: 25 November 2015), 2015.

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., Kindermann, G., Nakicenovic, N., and Rafaj, P.: RCP 8.5-A scenario

Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).



of comparatively high greenhouse gas emissions, Clim. Change, 109, 33–57, doi:10.1007/s10584-011-0149-y, 2011.

Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E. F., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R. W., Fabry, V. J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., and Foley, J. A.: A safe operating space for humanity, Nature, 461, 472–475, doi:10.1038/461472a, 2009.

Scheffer, M., Carpenter, S., Foley, J. A., Folke, C., and Walker, B.: Catastrophic shifts in ecosystems, Nature, 413, 591–596, doi:10.1038/35098000, 2001.

Schellnhuber, H. J.: Tipping elements in the Earth System, P. Natl. Acad. Sci., 106, 20561–20563, doi:10.1073/pnas.0911106106, 2009.

Shepherd, J., Caldeira, K., Cox, P., Haigh, J., Keith, D., Launder, B., Mace, G., McKerron, G., Pyle, J., Rayner, S., Redgewell, C., and Watson, A.: Working Group on Geoengineering the Climate, Geoengineering the climate: science, governance and uncertainty, London, UK, Royal Society, 98 pp., (RS Policy document, 10/29), 2009.

Smeets, E. M. W., Faaij, A. P. C., Lewandowski, I. M., and Turkenburg, W. C.: A bottom-up assessment and review of global bioenergy potentials to 2050, Prog. Energ. Combust., 33, 56–106, doi:10.1016/j.pecs.2006.08.001, 2007.

Snyder, P. K., Delire, C., and Foley, J. A.: Evaluating the influence of different vegetation biomes on the global climate, Clim. Dynam., 23, 279–302, doi:10.1007/s00382-004-0430-0, 2004.

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., Biggs, R., Carpenter, S. R., Vries, W. D., Wit, C. A. D., Folke, C., Gerten, D., Heinke, J., Mace, G. M., Persson, L. M., Ramanathan, V., Reyers, B., and Sörlin, S.: Planetary boundaries: Guiding human development on a changing planet, Science, 347, 1259855, doi:10.1126/science.1259855, 2015.

Sutton, R. T., Dong, B., and Gregory, J. M.: Land/sea warming ratio in response to climate change: IPCC AR4 model results and comparison with observations, Geophys. Res. Lett., 34, L02701, doi:10.1029/2006GL028164, 2007.

Tans, P. and Keeling, R.: Mauna Loa CO_2 annual mean data, ftp://aftp.cmdl.noaa.gov/products/trends/co2/co2_annmean_mlo.txt (last access: 2 February 2016), 2015.

Thomson, A. M., Calvin, K. V., Smith, S. J., Kyle, G. P., Volke, A., Patel, P., Delgado-Arias, S., Bond-Lamberty, B., Wise, M. A., Clarke, L. E., and Edmonds, J. A.: RCP4.5: a pathway for stabilization of radiative forcing by 2100, Clim. Change, 109, 77–94, doi:10.1007/s10584-011-0151-4, 2011.

UNFCCC: Adoption of the Paris Agreement, FCCC/CP/2015/L.9/Rev1, (United Nations Framework Convention on Climate Change), http://unfccc.int/resource/docs/2015/cop21/eng/10a01.pdf (last access: 3 September 2016), 2015.

van Kan, A., Jegminat, J., Donges, J. F., and Kurths, J.: Constrained basin stability for studying transient phenomena in dynamical systems, Phys. Rev. E, 93, 042205, doi:10.1103/PhysRevE.93.042205, 2016.

van Vuuren, D. P., Stehfest, E., Elzen, M. G. J. D., Kram, T., Vliet, J. V., Deetman, S., Isaac, M., Goldewijk, K. K., Hof, A., Beltran, A. M., Oostenrijk, R., and Ruijven, B. V.: RCP2.6: exploring the possibility to keep global mean temperature increase below 2 °C, Clim. Change, 109, 95–116, doi:10.1007/s10584-011-0152-3, 2011.

van Vuuren, D. P., Bayer, L. B., Chuwah, C., Ganzeveld, L., Hazeleger, W., Hurk, B. V. D., Noije, T. V., O'Neill, B., and Strengers, B. J.: A comprehensive view on climate change: coupling of earth system and integrated assessment models, Environ. Res. Lett., 7, 024012, doi:10.1088/1748-9326/7/2/024012, 2012.

van Vuuren, D. P., Lucas, P. L., Häyhä, T., Cornell, S. E., and Stafford-Smith, M.: Horses for courses: analytical tools to explore planetary boundaries, Earth Syst. Dynam., 7, 267–279, doi:10.5194/esd-7-267-2016, 2016.

Vaughan, N. E. and Lenton, T. M.: A review of climate geoengineering proposals, Clim. Change, 109, 745–790, doi:10.1007/s10584-011-0027-7, 2011.

Yuan, W., Luo, Y., Li, X., Liu, S., Yu, G., Zhou, T., Bahn, M., Black, A., Desai, A. R., Cescatti, A., Marcolla, B., Jacobs, C., Chen, J., Aurela, M., Bernhofer, C., Gielen, B., Bohrer, G., Cook, D. R., Dragoni, D., Dunn, A. L., Gianelle, D., Grünwald, T., Ibrom, A., Leclerc, M. Y., Lindroth, A., Liu, H., Marchesini, L. B., Montagnani, L., Pita, G., Rodeghiero, M., Rodrigues, A., Starr, G., and Stoy, P. C.: Redefinition and global estimation of basal ecosystem respiration rate, Global Biogeochem. Cy., 25, GB4002, doi:10.1029/2011GB004150, 2011.

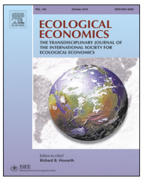
Reproduced from: V. Heck et al., Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal, Earth Syst. Dyn., vol. 7, no. 4, pp. 783-796, 2016, doi: 10.5194/esd-7-783-2016. Published under Creative Commons Attribution License 3.0 (CC BY).

Ecological Economics 159 (2019) 198–211

Contents lists available at ScienceDirect

Ecological Economics

journal homepage: www.elsevier.com/locate/ecolecon



Analysis

Can Intensification of Cattle Ranching Reduce Deforestation in the Amazon? Insights From an Agent-based Social-Ecological Model

Finn Müller-Hansen[a,b,*], Jobst Heitzig[a], Jonathan F. Donges[a,c], Manoel F. Cardoso[d], Eloi L. Dalla-Nora[e], Pedro Andrade[d], Jürgen Kurths[a,b], Kirsten Thonicke[a]

[a] *Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 60 12 03, D-14412 Potsdam, Germany*

[b] *Department of Physics, Humboldt University Berlin, Newtonstraße 15, Berlin 12489, Germany*

[c] *Stockholm Resilience Center, Stockholm University, Kräftriket 2B, Stockholm 114 19, Sweden*

[d] *Center for Earth System Science, National Institute for Space Research, Av. dos Astronautas 1758, São José dos Campos 12227-010, SP, Brazil*

[e] *Department of Research and Applied Technology, TerraSAT Company, Rua Pietro Cescon 1938, 99560-000 Sarandi, RS, Brazil*

ARTICLE INFO

Keywords:
Amazon deforestation
Land-use intensification
Pasture management
Social-ecological systems
Agent-based modeling

ABSTRACT

Deforestation in the Amazon with its vast consequences for the ecosystem and climate is largely related to subsequent land use for cattle ranching. In addition to conservation policies, proposals to reduce deforestation include measures to intensify cattle ranching. However, the effects of land-use intensification on deforestation are debated in the literature. This paper introduces the *abacra* model, a stylized agent-based model to study the interplay of deforestation and the intensification of cattle ranching in the Brazilian Amazon. The model combines social learning and ecological processes with market dynamics. In the model, agents adopt either an extensive or semi-intensive strategy of cattle ranching based on the success of their neighbors. They earn their income by selling cattle on a stylized market. We present a comprehensive analysis of the model with statistical methods and find that it produces highly non-linear transient outcomes in dependence on key parameters like the rate of social interaction and elasticity of the cattle price. We show that under many environmental and economic conditions, intensification does not reduce deforestation rates and sometimes even has a detrimental effect on deforestation. Anti-deforestation policies incentivizing fast intensification can only lower deforestation rates under conditions in which the local cattle market saturates.

1. Introduction

Can intensification of agricultural land use help us preserve threatened ecosystems such as the Amazon rain forest? If land is easily accessible, low-productivity land use often results in a high demand for land, putting pressure on ecologically important areas. Therefore, a common proposition is to increase yields per area to ease this pressure. In the economic literature, this proposition is often referred to as the Borlaug hypothesis (Angelsen and Kaimowitz, 2001, p. 3). The discussion mainly focuses on crop production, but livestock is equally important.

In the Amazon, livestock production, especially beef cattle ranching, drives expansion of pastures into the rainforest (Barona et al., 2010; Pacheco and Poccard-Chapuis, 2012). While more than 60% of the deforested area in the Brazilian legal Amazon was used as pasture by 2008, only about 5% was used for crop production (Almeida et al., 2016). In the last decades, the opening of the region for national and international markets has led to a shift from extractive land-use activities to cattle ranching and increased the activities of agribusiness including the development of a supply chain for meat processing (Salisbury and Schmink, 2007; Pacheco and Poccard-Chapuis, 2012). This increased the demand for agricultural land in the Amazon basin considerably, also via indirect effects (Richards et al., 2014). The expansion of pasture leads to large-scale deforestation with strong adverse impacts on biodiversity and local climate, for example, reduced precipitation as a result of lower evapotranspiration from deforested areas (Zemp et al., 2017). Lower precipitation in turn affects agricultural productivity (Oliveira et al., 2013) and may constitute a tipping element with relevance for global climate (Lenton et al., 2008).

On average, cattle ranching in the Amazon is characterized by extensive production systems with low stocking rates compared to other regions (Pacheco and Poccard-Chapuis, 2012). Many extensive

[*] Corresponding author at: Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 60 12 03, D-14412 Potsdam, Germany.

E-mail address: mhansen@pik-potsdam.de (F. Müller-Hansen).

https://doi.org/10.1016/j.ecolecon.2018.12.025
Received 25 March 2018; Received in revised form 21 November 2018; Accepted 29 December 2018

Reproduced from: F. Müller-Hansen et al., Can Intensification of Cattle Ranching Reduce Deforestation in the Amazon? Insights From an Agent-based Social-Ecological Model, Ecol. Econ., vol. 159, pp. 198-211, 2019, doi: 10.1016/j.ecolecon.2018.12.025, Copyright 2019 Elsevier B.V. All rights reserved, with permission from Elsevier.

production techniques can be linked to environmental degradation in the region. Slash-and-burn methods are used to fertilize the land and may spark unintended forest fires (Cano-Crespo et al., 2015). In many areas, nutrient-poor soils lead to fast run-down of pasture fertility (Serrão et al., 1979; Myers and Robbins, 1991). Additionally, weed invasion, pests, compaction, and erosion further promote pasture rundown (Landers, 2007). The exhausted pastures are often abandoned and secondary vegetation starts to regrow on them (Perz and Skole, 2003a,b). However, this forces the ranchers to replace them with pastures on newly deforested areas and move the frontier further into pristine forest.

Since the 2000s, there have been various efforts to reduce deforestation in the Brazilian Amazon (Nepstad et al., 2014). This includes the enforcement of environmental laws, which entails considerable costs and requires careful monitoring. As the current stagnation of deforestation rates shows, the present policy measures have their limitations (Azevedo et al., 2017). For example, Richards et al. (2017) show that agents react to the current monitoring system by deforesting smaller patches to avoid detection. Besides, current environmental legislation, the Brazilian Forest Code, allows land-owners to deforest 20% of their private lands (Soares-Filho et al., 2014). Cutting only the legally available areas will already lead to large losses in biodiversity and considerable amounts of greenhouse gases released into the atmosphere (Aguiar et al., 2016).

For these reasons, policies that promote the intensification of cattle ranching have been suggested as a viable option to reduce deforestation (Cohn et al., 2014). Intensification could help ranchers use the already deforested land more efficiently and detain them from deforesting more. These proposals are heavily criticized, arguing that higher profits from intensified land use may even increase deforestation rates (Angelsen and Kaimowitz, 1999; Kaimowitz and Angelsen, 2008). Other authors note that the success of intensification policies cannot be determined a priori but highly depends on the political, economic, and environmental circumstances (Latawiec et al., 2014).

Empirical evidence to support the effectiveness of intensification as a means to reduce deforestation in the Amazon is hard to assess and at most mixed. Cohn et al. (2011) review some of the cattle ranching intensification programs in Brazil that aim at the adoption of yield-increasing technology. They argue that due to a lack of data, the implementation of policies should proceed very carefully as it might result in unintended consequences. Soler et al. (2014) find that land-use developments in the federal states of Mato Grosso and Rondônia are strongly linked to market accessibility and the land distribution structure. They cannot uncover clear mechanisms that link land-use intensification to expansion of the deforestation frontier. Barretto et al. (2013) argue that land-use intensification in frontier regions coincides with the expansion of agriculture. An analysis of deforestation drivers also shows that intensified land use is associated with higher incomes, which in turn can be linked to higher deforestation (Busch and Ferretti-Gallon, 2017). After all, huge data gaps make the comparison of different management techniques of livestock systems difficult (Erb et al., 2016). A big challenge is to disentangle the effect of intensification from other influences and drivers (e.g., enforcement of legal protection) in empirical data. This also makes assessments of the impact of intensification policies difficult, mostly because of the huge heterogeneity of agents and their changing importance and roles in the deforestation process (Pacheco, 2012; Godar et al., 2014).

This paper investigates the interdependencies of intensification and deforestation using a theoretical modeling approach. Modeling has been used in the literature to investigate these interdependencies. For example, Bowman et al. (2012) use a spatial land rent model to find that intensification policies have to be complemented by improvements in conservation policies that disencourage land speculation to decrease deforestation. Many land-use models apply a procedure that determines demands for different types of land and then allocates them geographically. They use empirically derived statistics and economic criteria that indicate suitability of areas for different land uses. Conversion elasticities determine how changing demands translate into changes in spatial land-use patterns (e.g., Verburg et al., 2002; Michetti, 2012; Aguiar et al., 2012).

To intensify their production, ranchers have to adopt new management practices and production technologies. Such decisions are not only based on economic considerations, but are also determined by the diffusion of knowledge and successful management practices via social networks (Feder and Umali, 1993). This has been demonstrated and modeled for example for the adoption of new agricultural technologies (Berger, 2001; Maertens and Barrett, 2012). Therefore, it is important to consider the social and cultural context of cattle ranching intensification. For example, there are not only strong economic incentives but also cultural drivers, such as the dissemination and adoption of values that make the current practice of cattle ranching attractive in comparison with more sustainable land uses ("cowboy culture", Hoelle, 2011).

Agent-based approaches can capture such influences on land-use change. They model the decisions of heterogeneous agents and their social and environmental interactions to explain emergent patterns and dynamics at the system level. They can therefore describe how social interactions and incentive structures influence the decisions of ranchers to use the land in a specific way. Agent-based models (ABMs) are widely applied to describe social-ecological systems (for reviews see Schlüter et al., 2012; An, 2012; Groeneveld et al., 2017; Parker et al., 2003; Matthews et al., 2007; Heppenstall et al., 2012). In the land-use context, social-ecological ABMs are mostly developed for small study regions, taking into account local specificities and fitting behavioral patterns to data acquired in the field (Parker et al., 2008). There are several ABMs in the literature explicitly developed to study the influence of socio-economic drivers on deforestation dynamics. Many of these models use profit or utility maximization approaches to describe land-use decisions. For example, Andersen et al. (2017) provide a model of households in a small Bolivian community to explore the consequences of different policy options, including the level of public investment, a deforestation tax, and conservation payments. The model by West et al. (2018) is based on similar principles and focuses on the effects of direct REDD+ payments to agricultural households. Other models use heuristic approaches to land-use decisions, focusing for example on colonist households (Deadman et al., 2004) and on deforestation outcomes under different institutional settings (Costa, 2012) in frontier regions of the Brazilian Amazon. Some models also take local interactions between individual agents into account. For example, Mena et al. (2011) use socioeconomic surveys and demographic data to calibrate complex heuristic decision making modules in a model that describes households in the Ecuadorian Amazon. Manson and Evans (2007) combine different decision-making approaches in a genetic programming framework to model deforestation in Mexico. However, none of these models integrates social influence processes and their role for land-management decisions.

This paper presents the *abacra* (agent-based amazonian cattle ranching) model, a stylized ABM to investigate under which circumstances intensification of cattle ranching can reduce deforestation in Amazon frontier regions. The model described in Section 2 of this paper combines simplified representations of the social, economic, and ecological processes that we judge most important for the purpose of this study. It differs from the above-mentioned ABMs by specifying heuristic land-management strategies and capturing how these change as a result of social influence. Such a combination of approaches has been identified as a promising representation of human decision making in social-ecological models (e.g., Müller-Hansen et al., 2017). The model serves as a proof of concept that the combination of non-standard decision-making with local and social interactions can help to understand and explore the emergent system-level outcomes of social-ecological systems. It does not aim at producing concrete numerical predictions or scenarios of future land use in the Amazon.

After introducing the model, Section 3 provides a detailed analysis of the model results to demonstrate its dynamics, using data from the frontier region around Novo Progresso in southern Pará. Sections 4 and 5 discuss broader implications and limitations of the results and conclude the paper.

2. Model Description

In this section, we describe the details of the *abacra* model that we use throughout this study. A full description according to the ODD + D protocol (Müller et al., 2013) is provided in the Supplementary Material (see Appendix B).

2.1. Overview

The model is designed to investigate the interrelation between intensification of cattle ranching and deforestation in an Amazon frontier region. Furthermore, it demonstrates how social learning dynamics can be combined with heuristic land-management strategies and market dynamics to integrate social, economic and ecological dynamics. The model is designed for researchers interested in tropical deforestation, land modeling and complex social-ecological systems.

The model comprises a large number N of ranchers with their respective land properties. The ranchers interact with their local environment by decisions to convert forest into pasture and managing this pasture. The land area of every ranch is divided into three different land-cover categories (forest, pasture, secondary vegetation). Furthermore, the pasture productivity and the soil quality of areas with secondary vegetation describe the environmental quality of the land. Land-cover succession equations trace deforestation, land abandonment, and forest regrowth, while two other dynamic equations describe the evolution of the productivity of pasture and secondary vegetation.

The ranchers are characterized by their savings and their land-management strategy. The decisions of agents are captured by heuristic strategies depending on economic and ecological constraints. Agents can follow either an extensive strategy, corresponding to traditional cattle ranching with fallow periods and slash-and-burn fertilization, or a semi-intensive strategy. In contrast to intensive cattle ranching that relies mostly on externally produced feedstock, semi-intensive cattle ranching increases the productivity of the pasture on which the cattle graze by inputs such as machinery and fertilizers. The choice of the land-management strategy is modeled as a social learning process: Agents are located on a geographic network representing neighborhood and acquaintance relations. They imitate the successful strategies of their neighbors. Key parameters of the model describe the cattle market demand and the time scale of social learning.

The model is discrete in time t and each time step represents one year, thereby abstracting from seasonal variations. The simulation for each time step proceeds in the following sequence:

First, the agents make decisions about their land-use activities, based on the previous state of their environment and their economic situation (Sections 2.4–2.6). Second, based on the previous state and the decisions, the system evolves according to the environmental dynamics (Section 2.2). Third, all ranchers receive revenues for the cattle they produced (Sections 2.3 and 2.8). Finally, ranchers imitate their neighbors' land-management strategies with a probability depending on the difference of the rancher's consumption with its neighbor (Section 2.7).

The model is implemented in python, using various packages of the python ecosystem (numpy, scipy, pandas, networkx) to combine the data with the dynamics described above.[1] This language was chosen to allow an easy parallelization of model runs on the high performance cluster computing infrastructure of the Potsdam Institute for Climate Impact Research.

In the following, we describe the different parts of the model in detail. Table 2 in Appendix A gives an overview of the variables used for the formalization.

2.2. Ecological Dynamics

Each agent i has a ranch with a constant area X that is covered by forest F_t, pasture P_t, and secondary vegetation S_t. Thus, $F_t + P_t + S_t = X$, where we drop the index i indicating the rancher. Land-cover changes such as deforestation and land abandonment are traced by land-cover succession equations (cp., e.g., Satake and Rudel, 2007). At each time step, pasture land can be created through deforestation d_t or reuse of land previously covered by secondary vegetation r_t. Pasture with area a_t can also be abandoned, leading to secondary vegetation regrowth. The change in pasture land is given by

$$P_{t+1} = P_t + d_t + r_t - a_t, \tag{1}$$

where d_t, r_t, and a_t are rates per year in units of area. The dynamics of forest and secondary vegetation are given by

$$F_{t+1} = F_t + r_n v_t S_t - d_t \tag{2}$$

and

$$\begin{aligned} S_{t+1} &= X - P_{t+1} - F_{t+1} \\ &= S_t - r_n v_t S_t + a_t - r_t, \end{aligned} \tag{3}$$

where r_n is a parameter that describes the natural recovery from secondary vegetation to mature forest. The deforestation d_t, abandonment a_t, and reuse r_t are control variables determined in the rancher's decision process. The land-cover dynamics for a single ranch are illustrated in Fig. 1.

The pasture land is furthermore characterized by an average productivity q_t. The agent can decide how much cattle to place on the pasture. Pasture productivity is decreasing if the stocking rate $l_t = L_t/P_t$ is high, i.e., there is a high number of cattle L_t per area on the pasture. The model formulation implicitly assumes here that the herd size of ranchers is variable through acquisition and sale of calves and the ranchers adjust it to their requirements (cp. Quaas et al., 2007). The decay of pasture productivity can be reduced by a management effort m_t, which subsumes various processes like fertilization, adoption of new grass species, fencing, and maintenance work.

For describing the dynamics of the pasture productivity, we choose the simplest decreasing dynamics with a lower zero bound, i.e., an exponential decay. Deforestation and reuse add land area to the pasture with productivities q_d and v_t, respectively. Furthermore, abandonment lets the pasture area shrink. Averaging over all these changes and weighting with the respective areas gives the following dynamics for pasture productivity:

$$q_{t+1} = \frac{(1 - \beta(l_t - m_t))q_t(P_t - a_t) + q_d d_t + v_t r_t}{P_t + d_t + r_t - a_t}, \tag{4}$$

where β is the rate of degradation, l_t is the stocking rate of the pasture, and m_t is a management effort that can counteract pasture degradation.

Finally, the variable v_t tracks the productivity and regrowth on land areas with secondary vegetation. It follows a similar dynamics as the pasture productivity, but with an exponential approach to the natural relative productivity $v^* = 1$ with rate r_S. The other terms stem from weighting and averaging for additional and outgoing areas, similar to Eq. (4).

$$v_{t+1} = \frac{(v_t + r_S(1 - v_t))(S_t - r_t) + a_t q_t}{S_t - r_t + a_t}. \tag{5}$$

In summary, the ecological state of each ranch has four degrees of freedom (P_t, F_t, q_t, and v_t).

[1] The code is available from www.github.com/fmhansen/abacra.

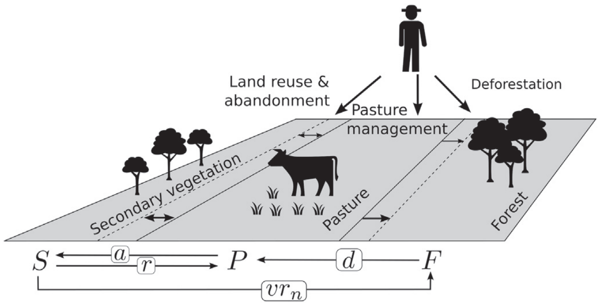


Fig. 1. Illustration of the conversion of land for single ranches in the model. The total area of a property is divided into three land-cover types that can be converted by land management with rates *d* (deforestation), *a* (abandonment), and *r* (reuse). Secondary vegetation regenerates with a rate proportional to a natural recovery parameter r_n and the productivity of secondary vegetation *v*. Cattle raised on the pasture generate revenues for the rancher.

2.3. Economic Dynamics

There are five control variables of the ecological dynamics, representing the possible decisions for the rancher: The management m_t, deforestation d_t and reuse r_t are associated with a cost per area. The income of the agent is realized from selling cattle $y_t = l_t P_t q_t / T_p$ at a price of p_c (per head), where T_p is the average time that cattle have to spend on the pasture until they can be slaughtered. Thus, the income of the agent is given by

$$I_t = p_c l_t P_t q_t / T_p - c_D d_t - c_R r_t - c_m m_t P_t, \tag{6}$$

where c_D and c_R are the cost of deforestation and reuse (per area) and c_m the cost of management (per area and effort).

This income can either be consumed or saved by the rancher, resulting in the following dynamics for the accumulated savings:

$$k_{t+1} = (1 + \delta)k_t + I_t - C_t, \tag{7}$$

with an interest rate δ. The income spent for consumption C_t also comprises a control in the model. Note that the savings can also be negative, such that they effectively represent the debt of the rancher. For reasons of simplicity, we assume here a fixed saving rate *s*, such that $C_t = (1 - s)I_t$.

2.4. Decision Making of Agents and Land-Management Strategies

The decision-making functions of agents are the centerpiece of the *abacra* model. They determine the amount of deforestation, abandonment, reuse, stocking rate, and pasture management in every time step. Because the land-use decisions may depend on many factors such as location, available resources, weather, beliefs about future prices and policies, and the choices of other agents, it is especially challenging to capture them appropriately in a stylized model.

Here, we use a heuristic decision approach for modeling the decisions of the ranchers. Heuristics are rules of thumb, often formalized as decision trees, that help agents to evaluate available information and choose actions that lead to more desirable outcomes over less desirable ones (for a recent review, see Gigerenzer and Gaissmaier, 2011).

As evidence from surveys suggests, land use decisions are not only based on monetary incentives but strongly influenced by social preferences (Garrett et al., 2017). Because of limited empirical data on actual decision processes in the system under consideration, we make the following simplifying assumptions for the agents' decision functions. We capture the social aspects of land-use decisions in our model by a heuristic land-management strategy that an agent adopts. This strategy determines how an agent makes use of the land. In the model, we implement two idealized strategies, an extensive and a semi-intensive land management strategy. They correspond to typical individual land-use trajectories in the Amazon.

2.5. Extensive Strategy

The extensive strategy represents traditional approaches to cattle ranching with fallow periods and slash-and-burn fertilization. It is characterized by low stocking densities. The pasture productivity decreases over time and has to be renewed by fallow periods and slash-and-burn practices.

The decisions to deforest or reuse (i.e., slash-and-burn) an area *D* or *R* are determined as follows. First, the respective savings for covering the conversion costs c_D or c_R have to be available. The conversion can only take place, if there is enough forest F_t or secondary vegetation S_t. For the extensive strategy, the managed pasture cannot exceed a fixed fraction of the total area p_{max} because the rest is set aside as fallow land. Finally, the expected additional income $I^d_{exp} = p_c l_t D q_d / T_p$ (or $I^r_{exp} = p_c l_t R v_t / T_p$ for reuse) from the additional pasture is compared to the cost. If the investment is paying back within a time period T_{rec}, the investment is made. If both deforestation and reuse are profitable, then the option with the higher expected additional income is chosen. The latter depends on the expected cattle price times the expected amount of cattle that can be produced on the new pasture. An area *A* of land is abandoned if pasture productivity falls below a certain threshold $q_{\theta a}$ and this land was used as pasture before.

The extensive strategy does not use the pasture management option ($m_t = 0$) and the stocking rate is fixed at a low level $l_t = l_{ext}$. The logic of the decisions are illustrated as two decision trees in Fig. 4. For the implementation of the model, we used Heaviside step-functions. The equations are given in the Supplementary material.

Fig. 2 shows a sample trajectory of a single ranch with the extensive strategy. The strong oscillations in the trajectory result from the thresholds in the decision functions. The agent has to reinvest into deforestation and reuse of secondary vegetation in order to improve the pasture productivity every few years.

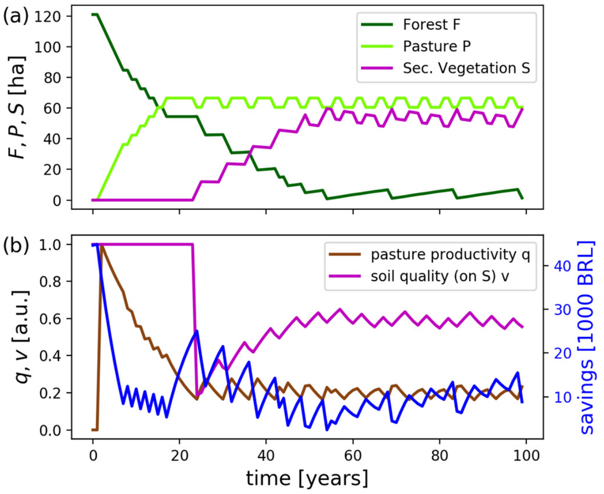


Fig. 2. Sample trajectory for illustration of the dynamics of a single ranch with extensive strategy, showing (a) the areas of different land use: pasture (light green), forest (dark green), and secondary vegetation (magenta) and (b) savings (blue), pasture productivity (brown), and secondary vegetation fertility (magenta), which are displayed in arbitrary units (a.u.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.6. Semi-intensive Strategy

The semi-intensive strategy, corresponding to cattle ranching with various industrial inputs and pasture improvement techniques, has higher stocking densities but also higher costs for inputs. Agents invest in inputs for pasture maintenance such as fertilizers and fencing for pasture rotation, but also in measures such as better adapted grass and cattle species, improved pasture seeding with legumes, or additional concentrated feed to improve pasture and livestock productivity.

The semi-intensive strategy is implemented in the following way: Deforestation D occurs if there is enough primary forest on the property left and the agent has sufficient savings to cover the deforestation cost. Furthermore, the agent evaluates whether it is possible to recover the investment within a certain time period T_{rec}, assuming that the economic circumstances remain constant: The agent compares the expected income $I^d_{exp} = p_c l_t D q_d / T_p - c_m m_t D$ from using a newly deforested area to the deforestation cost. The agent uses a similar logic to determine whether it is profitable to convert an area of secondary vegetation R back to pasture. As for the extensive strategy, the decision between deforestation or reuse to get new pasture results from a comparison of the expected income increases of both options. An area A of pasture is abandoned if the ranching activity is not profitable anymore.

For the semi-intensive strategy, the deforestation costs are higher by the intensification cost c_I. This also has to be considered in Eq. (6) by subtracting the intensification cost $c_I(d_t + r_t)$ for converted areas. Similarly, when adopting this strategy, the cost for converting existing pasture $c_I P_t$ has to be subtracted from the savings stock, Eq. (7). A formulation of these rules in terms of Heaviside functions is provided in the Supplementary materials.

The semi-intensive strategy uses the pasture management option $m_t = M$, where M is a constant. The stocking rate is higher than in the extensive case $l_t = l_{int} > l_{ext}$. A sample trajectory for this strategy is shown in Fig. 3. Here, one can observe that most of the forest is deforested quite fast and the decline of pasture productivity is much slower because of pasture management.

Evidence for the proposed kind of heuristic behavior was obtained in personal interviews by one of the co-authors (E. D.-N., unpublished fieldwork carried out in 2016 in the states of Pará and Mato Grosso along the highway BR-163). Ranchers tend to invest in new pasture if they can recover their initial investment in a time period below a threshold of about 5–8 years. Furthermore, the valuation of land is an important factor for decision making of ranchers. Because our model does not contain a description of the land market, we do not consider this in our analysis.

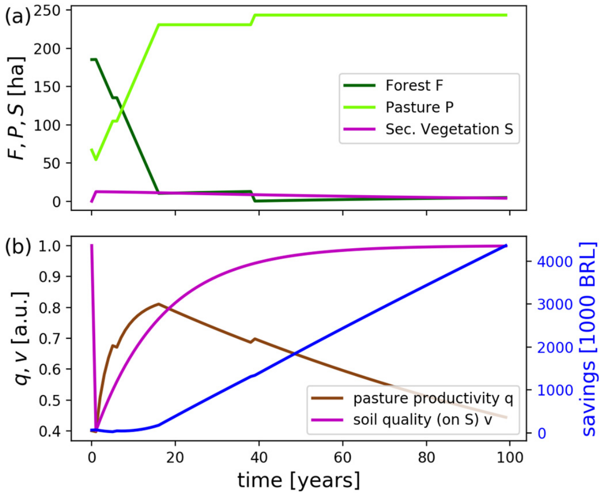


Fig. 3. Sample trajectory for illustration of the dynamics of a single ranch with the semi-intensive strategy: (a) areas of different land use: pasture (light green), forest (dark green), and secondary vegetation (magenta). (b) Savings (blue), pasture productivity (brown) and secondary vegetation fertility (magenta), which are displayed in arbitrary units (a.u.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

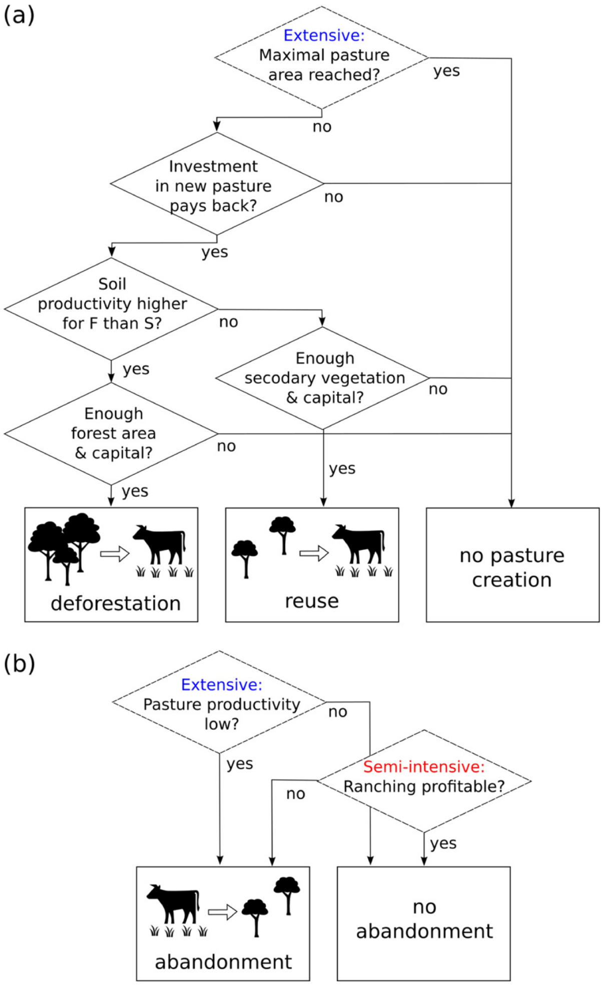

(a)
Extensive:
Maximal pasture area reached?
yes
no
Investment in new pasture pays back?
no
yes
Soil productivity higher for F than S?
no
yes
Enough secodary vegetation & capital?
no
yes
Enough forest area & capital?
no
yes
deforestation
reuse
no pasture creation
(b)
Extensive:
Pasture productivity low?
no
yes
Semi-intensive:
Ranching profitable?
no
yes
abandonment
no abandonment

Fig. 4. Decision trees illustrating the decision heuristics used by the agents in the model (a) for deforestation and reuse and (b) for abandonment. Differences between the extensive and the semi-intensive strategy are marked as dashed boxes. The differences regarding the stocking rate and the use of pasture management are not displayed.

2.7. Local Interaction: Strategy Imitation Between Agents

In the *abacra* model, we reduce the potentially complex process of adopting a land-management strategy to a social imitation process on a geographic network and assume that the adoption of a certain management strategy only depends on the agent's own success and its comparison with the neighbors (cp. Traulsen et al., 2010; Wiedermann et al., 2015). The agents are modeled on a network that represents neighbor relations as illustrated in Fig. 5. This simplifying assumption is motivated by evidence from the literature that neighbor interactions play an important role in deforestation decisions (Robalino and Pfaff, 2012) and the role of networked social interactions in various environmental contexts (Currarini et al., 2016). Furthermore, word-of-

Reproduced from: F. Müller-Hansen et al., Can Intensification of Cattle Ranching Reduce Deforestation in the Amazon? Insights From an Agent-based Social-Ecological Model, Ecol. Econ., vol. 159, pp. 198-211, 2019, doi: 10.1016/j.ecolecon.2018.12.025, Copyright 2019 Elsevier B.V. All rights reserved, with permission from Elsevier.

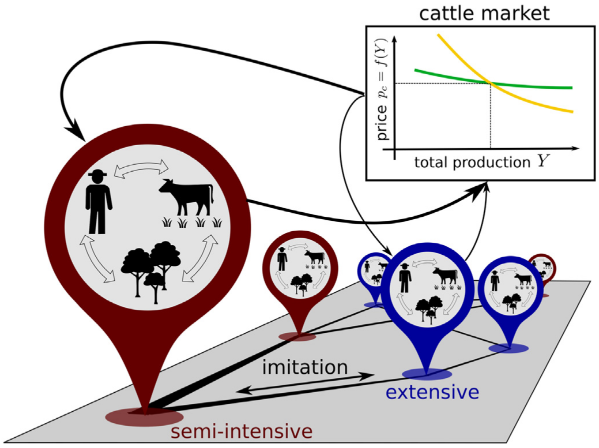


Fig. 5. Illustration of the local and system-wide interactions between agents: Agents can imitate their strategies (extensive, blue, or semi-intensive, red) if they are connected on a geographically embedded social network. They sell their cattle on a market that determines the cattle price and thus their income, depending on the price elasticity of demand (yellow curve: low price elasticity, green curve: high price elasticity). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

mouth recommendation has been identified as one of the most important determinants for the participation in sustainable ranching programs (zu Ermgassen et al., 2018).

We implement the neighbor interactions as follows: The simplest assumption for the timing of interaction events is that they are equally probable for every point in time, i.e., they occur with a constant imitation rate λ. Such a stochastic process is called Poisson process and is described by a rate λ (Van Kampen, 2007). The number of interaction events K in one time step of the model is then given by a random number drawn from a Poisson distribution with rate λ. For each interaction event, a random node i of the network and a random neighbor j of this node are chosen. Then, i imitates the strategy of j with a probability given by a hyperbolic tangent function of the difference between the agents' consumption C_t (cp. Wiedermann et al., 2015):

$$P_{ij} = \frac{1}{2}(\tanh(C_j - C_i) + 1). \tag{8}$$

However, the imitation of the intensive strategy is only possible if an intensification cost per area c_I can be covered. This cost can also be payed by a credit (modeled as negative savings) up to a certain limit k_{min}. The imitation process results in a faster spread of production strategies that generate more income.

2.8. Interaction Between All Agents: The Cattle Market

Additionally to the local imitation, the model captures how ranchers interact on a cattle market, which determines the price that ranchers can realize when selling their cattle. We model the price as given by a demand curve that represents the demand side of a local market for cattle. The price response to changes in cattle quantity $Y = \sum_i q_i P_i l_i$ is modeled by a constant elasticity function

$$p_c = a_p Y^{-1/\varepsilon}, \tag{9}$$

with price elasticity of demand ε. A high price elasticity means that a slight change in price leads to a high change in demand. Put the other way around, a large change in the quantity leads to a slight change in price. A low elasticity thus implies a strong price reaction to a change in the produced quantity. This relation is illustrated in the upper right of Fig. 5, where the yellow demand curve corresponds to a lower elasticity than the green one.

The price elasticity allows modeling different market settings: The

Table 1
Description, symbols, and values of parameters in the presented ABM. Where applicable, ranges in the literature for the parameterization with the corresponding sources or own calculations are indicated.

Parameter	Symbol	Default value	Range	Unit	Sources and comments
Deforestation cost	c_D	1500	1000–3000	BRL/ha[a]	Difference in land prices between pasture and Forest from FGVIBRE (n.d.)
Reuse cost	c_R	500	500–2000	BRL/ha	
Pasture maintenance cost	c_m	150	150–300	BRL/ha	Estimated using IMEA (n.d.)
Intensification cost	c_I	500	300–1000	BRL/ha	zu Ermgassen et al. (2018)
Live cattle price		5	3.4–6.4	BRL/kg	SEAB (n.d.)
Slaughter age	T_p	3	2.5–3	Years	Tab. 4 in Pacheco and Poccard-Chapuis (2012)
Cattle weight at slaughter (3 years)		500	470–520	kg	Tab. 4 in Pacheco and Poccard-Chapuis (2012)
Initial live cattle price	$p_c(0)$	2500	1600–3330	BRL/head	Live cattle price × weight at slaughter
Average stocking rate	l_{ext}, l_{int}	0.8, 1.6	0.5–2.0	Head/ha	Tab. 3 & 4 in Pacheco and Poccard-Chapuis (2012)
Saving rate	s	0.25	0.15–0.3		Gross domestic savings (The World Bank, n.d.)
Natural recovery parameter	r_n	0.013		1/year	Corresponding to a half-life of about 50 years (Poorter et al., 2016)
Regeneration of soil quality of Secondary vegetation	r_S	0.06		1/year	Corresponding to a half-life of about 10 years (Davidson et al., 2007)
Parameter of pasture degradation	β	0.15		1/head/year	Corresponding to a half-life of 3–4 years for Degradation (Costa, 2012)
Productivity of pasture after Deforestation	q_d	1		Arbitrary units (a.u.)	Determines scale
Threshold on q for abandonment	$q_{\theta a}$	0.2		a.u.	
Relative deforested, abandoned and Reused areas	D/X, R/X, A/X	0.05	0.02–0.1	Relative area	For deforestation, estimations with PRODES (n.d.) yield 0.08
Maximum relative pasture for Extensive strategy	p_{max}	0.5		Relative area	
Time period for investment decisions	T_{rec}	7		Years	Information from personal interviews: 5–8 years
Management effort	M	1.5		a.u.	
Maximal credit for intensification	k_{min}	200		BRL/ha	
Imitation rate	λ	1	0.001–10	1/year	
Price elasticity of demand	ε	10	0.1–1000		
Share of teleconnections	α	0.02	0–0.1		

[a] Prices are in 2010 Brazilian Real (BRL), areas are in hectare (10,000 m^2).

Reproduced from: F. Müller-Hansen et al., Can Intensification of Cattle Ranching Reduce Deforestation in the Amazon? Insights From an Agent-based Social-Ecological Model, Ecol. Econ., vol. 159, pp. 198-211, 2019, doi: 10.1016/j.ecolecon.2018.12.025, Copyright 2019 Elsevier B.V. All rights reserved, with permission from Elsevier.

price elasticity is lower and thus prices are more sensitive to changes in quantity in regions with a market that is not well integrated into national or international markets. If markets are well connected to bigger markets, the prices will not be affected much by changes in locally produced quantities but rather by external price fluctuations. The special case of fixed prices (ranchers being price takers) is effectively equivalent to very high price elasticities: in this case, the exponent in Eq. (9) gets close to zero such that the dependence on Y becomes negligible and the curve approaches the constant a_p. Instead of studying the case of fixed prices separately, we will look at very high values for the price elasticity.

2.9. Input Data and Parametrization

We use different data sources to estimate parameters of the *abacra* model. The details are given in Table 1. Some of the parameters, especially those related to decision making, cannot be determined from the data. We analyze the sensitivity of model outcomes on such input parameters further in Section 3.2.

The model uses the following input data for initial conditions and set-up of the network: Initial values for pasture areas are approximated from deforestation data from PRODES, using the data from 2000 as initial conditions. For comparison with other initial conditions, we also test initial conditions corresponding to the deforestation extent in 2016. The initial conditions for secondary vegetation are set to zero. Initial values for the soil productivity q are randomly drawn from a uniform distribution of values between 0 and 1. Furthermore, we allocate initial savings to the ranchers drawn from a log-normal distribution with mean 200 and standard deviation 100 BRL per ha of property area.

We apply the model framework on the study region around Novo Progresso in the Brazilian Amazon. We choose the region because it is characterized by strong deforestation in recent years and a high share of cattle ranching on deforested areas. However, the model should be easily adapted to other regions and could be scaled to larger regions.

We use property data from the Rural Environmental Registry (Cadastro Ambiental Rural, CAR), a geoinformation tool that helps the administration to monitor land owners' compliance with the Forest Code (Azevedo et al., 2017). Between 2000 and 2016, average deforestation on CAR-registered properties in the Novo Progresso region was 9.5 ha/year with an average property size of 563 ha. In total, about 28% of the forest area on registered properties has been cleared by 2016 (own calculations using PRODES and CAR).

We use the CAR data to get a representative heterogeneity of property sizes and construct different neighborhood networks. However, the CAR data is incomplete and contains unsettled land claims, which leads to overlapping properties. To avoid inconsistencies, we remove properties with large overlap by via visual inspection of the data set in a GIS program. Like this, we remove properties that overlap with more than a small part of their total area. Fig. 6 (a) shows the municipality of Novo Progresso and its adjacent municipalities as well as the limits of properties in the CAR data.

To construct the network, we apply a function on the distance between properties (nodes) determining whether they are connected or not. The simplest method connects all properties closer than a specific threshold. We test the model with networks for different thresholds and choose 10 km because this results in a good balance between overall connectivity of the network and an average degree that is in a reasonable range for social contacts. This network has 4012 nodes and an average degree of about 81.

We also test probabilistic methods for constructing neighborhood networks, for which the probability of being connected decays exponentially with the distance between properties. Furthermore, we construct geographic networks that have a proportion α of links replaced by random links. We call these links teleconnections because they are independent of the spatial embedding of the network and therefore represent social interactions over distance. Fig. 6 (b) shows the network constructed from the property data without teleconnections. For the model simulations, the initial strategies are set as follows: all properties start with the extensive strategy except the ones within a range of 10 km from the major cities, which start with a 50% probability with the semi-intensive strategy. The colors of the network nodes in the figure indicate initial conditions for the agents' strategies.

3. Model Analysis and Results

After introducing the model design in the previous section, this section discusses system-level outcomes of model simulations with interacting agents.

3.1. System-Level Dynamics

For parameter settings with a high imitation rate λ and high elasticity of demand ε, the initially small number of agents with a semi-intensive strategy increases over time until almost all agents use this strategy. This happens because the increase in produced cattle does not decrease the revenue per area significantly. Further deforestation allows more cattle to be raised and thus increases overall income, which can be reinvested to deforest more.

Fig. 7 shows the key variables of an ensemble of model runs with such a parameter setting (the other parameters are given in Table 1). The shaded ranges indicate the variation of variables due to different realizations of the stochastic processes in the model. The figure shows that most of the forest is already deforested and converted to pasture in the first 30 to 40 years of the simulation (panel a). Panel (b) in Fig. 7 shows that after an initial peak in pasture productivity stemming from newly deforested pastures with a high initial productivity, q drops because of ongoing pasture degradation. Later, it increases as more and more agents use pasture management to improve their pasture productivity. The productivity of secondary vegetation is initially low, but increases as the soil regenerates. The agents' savings are low at the beginning and accumulate at the end of the simulation as many agents have already deforested all of their area and cannot invest in more pasture. The fraction of ranchers that adopted the semi-intensive strategy in panel (c) of Fig. 7 increases rapidly, because they have the possibility to borrow money for intensification. In a scenarios in which this option is not available, they first have to accumulate the savings to cover intensification costs, which slows down the increase. For higher imitation rates and higher cattle prices, this fraction increases more rapidly. Panel (d) in Fig. 7 finally shows how the produced cattle quantity Y increases rapidly in the first 40 years. After all forest has been converted to pasture, there is a slow decay due to pasture degradation. The cattle price p_c hardly changes because of the high price elasticity.

For comparison, Fig. 8 displays the results of model simulations with similar parameterization except for a lower imitation rate and lower elasticity. Here, one can observe that because of the low imitation rate, the number of ranchers with a semi-intensive strategy increases only slowly (Fig. 8 (c)). This leads to the abandonment of degraded pasture and an increase in secondary vegetation (Fig. 8 (a)). Furthermore, the low price elasticity of demand leads to a strong reaction of prices to increasing production at the beginning of the simulation, as a comparison of Figs. 7 and 8 in panels (d) illustrates. As the pastures degrade and production goes down, the price recovers towards the middle of the displayed simulation time. At the end of the simulation, prices decrease again because intensification sets in and cattle production increases. In the long run, the lower revenues lead to less savings (Fig. 8 (b)) and thus slow down deforestation, as panel (a) in Fig. 8 illustrates.

A formal analysis of the asymptotic dynamics of the model is difficult because the system is very heterogeneous and stochastic. Long-term simulation results suggest that there are (quasi) stable states and cyclic asymptotic dynamics, depending on the parameter regime. They

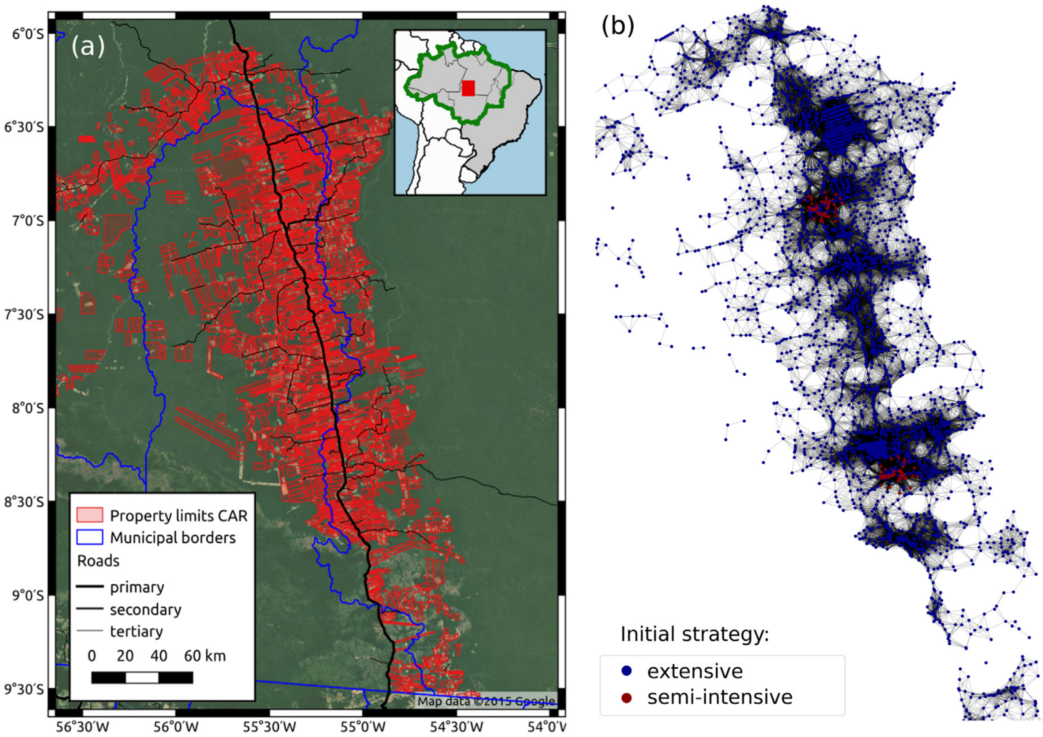


Fig. 6. (a) Map of the study region with property limits from the environmental registry (CAR; red), municipality borders (blue) and roads (black). The data is plotted over a satellite image of the region. The inset shows the location in Brazil (grey) and the Brazilian legal Amazon (green line). (b) Geographic neighborhood network without teleconnections ($\alpha = 0$) derived from this data. Each node represents a property. The color of the nodes depicts the distribution of initial strategies. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

are only reached after long transients (several hundred years) as an effect of the slow forest recovery dynamics. We do not analyze them in detail, because we are interested in deforestation, which is mainly a transient phenomenon.

3.2. Sensitivity Analysis

Here, we present an analysis of how model results depend on specific model parameters. Several parameters are difficult to estimate due to a lack of data and therefore a sensitivity analysis is crucial. Parameters may also change over time and an analysis of the dependence of model outcomes can illustrate how trends in external drivers of the system might influence model outcomes. We focus our analysis on six parameters describing costs and prices as well as the imitation process. An exploration of further results indicates that variations of other parameters do not lead to qualitatively different model behavior.

Price elasticity of demand and deforestation cost are crucial for the revenues and production costs of ranchers. They have a direct influence on the production of cattle and the rate of deforestation. A lower elasticity inhibits the expansion of cattle production and deforestation (Fig. 9), while higher deforestation costs slow down deforestation (see Fig. S1 in the Supplementary Material). The former is due to a saturation of the local cattle market. The effect of both parameters on intensification is limited.

The four parameters imitation rate, intensification cost, limitations to intensification credit, and teleconnection share influence the imitation of strategies and therefore directly impact the speed of the spread of the semi-intensive pasture management strategy. Fig. 10 shows how a lower imitation rate leads to a considerably slower spread of the semi-intensive strategy. This also leads to a lower cattle production and deforestation. A higher intensification cost inhibits fast intensification and thus the expansion of pasture and cattle production (Fig. S2). The same applies for low limits to credit that a rancher can access (parameter k_{min}). If ranchers cannot access credit at all ($k_{min} = 0$), the intensification process is considerably slowed down (Fig. S3). Finally, the share of teleconnections has only limited influence on the speed of intensification. However, if we do not add teleconnections to the network of neighboring ranches, some of the ranches are isolated. Therefore, they cannot adopt the semi-intensive strategy at all. This leads to a saturation of the intensification share below 1 (Fig. S4).

To make it more systematic, we extended the analysis to aggregate measures of the transient model behavior. Because this study analyzes the interaction between intensification and deforestation, we focus on the impact of different parameter combinations on the average deforestation. Figs. 11 and 12 show the mean over the first 50 years after model initialization, because this is the period in which most of the deforestation happens (compare Figs. 7 and 8).

In Fig. 11, the average deforestation is plotted depending on the elasticity of the cattle demand function as well as the imitation rate (both on a log-scale). The results match with observed mean deforestation rates on properties ranging between 3 and 20 ha/year (own calculations using PRODES and CAR). The figure shows that for low imitation rates and elasticities, the average deforestation is in the medium range of 3–4 ha/year. For low elasticity, this decreases with a higher imitation rate, which is associated to faster intensification. For a high elasticity of demand, this relationship is reversed: A higher imitation rate increases the higher deforestation rate even further.

If there are high intensification costs and agents do not have access to credit, the intensification under high imitation rates is hampered. Therefore, such conditions will not result in an increase of deforestation under high imitation rates (see Fig. S5).

We also test other parameter ranges indicated in Table 1 and find

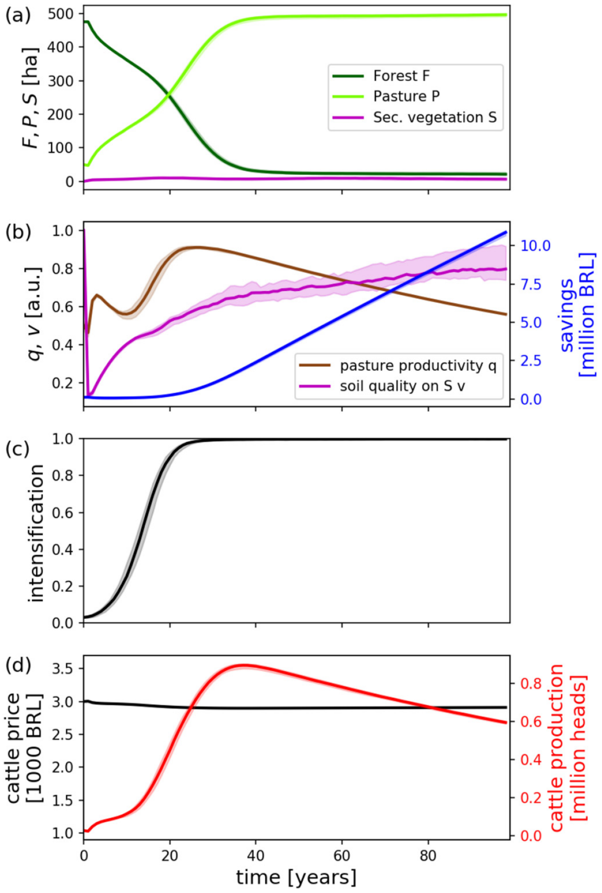


Fig. 7. Mean state variables of agents on the geographic network depicted in Fig. 6 with high imitation rate (λ = 1), high elasticity ($\varepsilon = 100$), and some teleconnections in the social network ($\alpha = 0.02$): (a) mean areas (forest, pasture, secondary vegetation), (b) mean pasture productivity, soil quality on secondary vegetation areas, and savings, (c) intensification: ratio of ranches with the semi-intensive strategy (red nodes in Fig. 6), and (d) price and quantity of produced cattle. The thick lines are the respective ensemble median of a sample of 1000 model runs with different realizations of the stochastic processes in the model and the shaded areas around them indicate the 5th to 95th percentile of the distribution of model outcomes. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

that even though they may influence the results quantitatively, they do not change the model outcomes in a relevant way. For instance, variation of the parameter determining the relative areas that agents can deforest preserves our main findings (see Fig. S6).

The results presented here are properties of the transient dynamics of the system, not equilibrium or asymptotic states. Therefore, they depend on the initial conditions of the system, especially on the initial pasture areas, pasture productivity, and savings. We test the dynamics for different settings of initial conditions and find for all of them that an increase in imitation rate does not reduce deforestation rates if the price elasticity is high.

3.3. Network Effects

Apart from the influence that certain parameters and initial conditions have on the model outcome, we also investigate the influence of the topology of the underlying neighborhood network. To account for long-range social ties (i.e., family and friendship relations independent of geographic distance), we test how the spreading of land-management strategies on the social network changes if we replace a fraction of local links by teleconnections, i.e., random links that are independent of the spatial embedding (cp. Section 2.9).

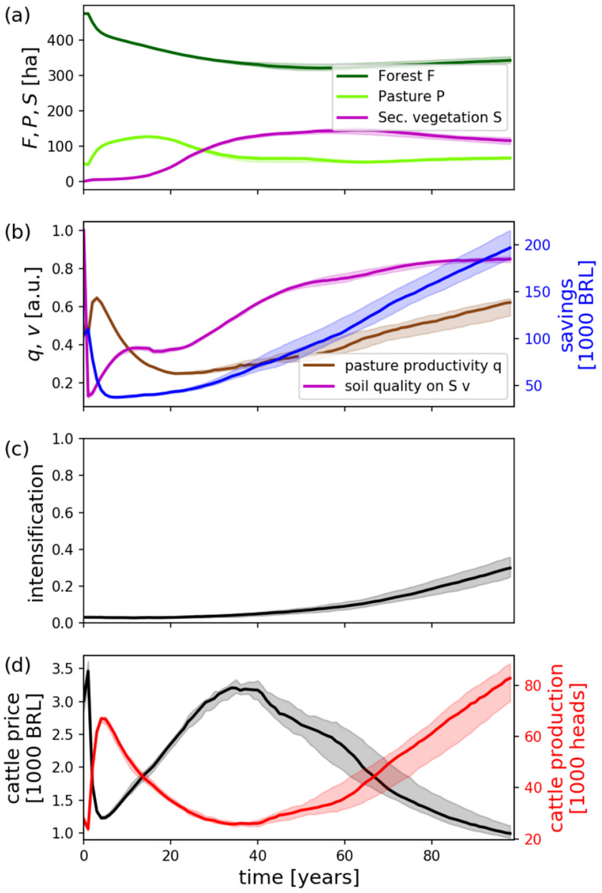


Fig. 8. Mean state variables of agents on the geographic network (Fig. 6) but with lower imitation rate (λ = 0.1) and elasticity ($\varepsilon = 1$). The shown variables are the same as in Fig. 7.

For random initial conditions with a spatially uniform distribution, the spreading does not change strongly when replacing a fraction α of local connections with teleconnections. With initial conditions for which ranches with semi-intensive strategies are spatially concentrated (e.g., around local cities or main roads), the additional teleconnections accelerate the spreading of the strategies considerably. Under parameter settings where the semi-intensive strategy is favored, the intensification process is therefore accelerated by the introduction of teleconnections.

Fig. 12 displays the average deforestation rate depending on the share of teleconnections in the network and the imitation rate. For medium imitation rates, the influence of the teleconnection share on the deforestation outcome is small compared to other effects in the model. The figure suggests that adding teleconnections has the same effect as slightly rescaling the imitation rate.

In addition to the network construction as described in Section 2.9, we also test a method for network construction that links nodes with a probability that decays exponentially with distance (Waxman, 1988). This results in changes in the network structure because the threshold on the distance is replaced by a characteristic length for the decay. It has only very limited effect on model outcomes and does not change

Reproduced from: F. Müller-Hansen et al., Can Intensification of Cattle Ranching Reduce Deforestation in the Amazon? Insights From an Agent-based Social-Ecological Model, Ecol. Econ., vol. 159, pp. 198-211, 2019, doi: 10.1016/j.ecolecon.2018.12.025, Copyright 2019 Elsevier B.V. All rights reserved, with permission from Elsevier.

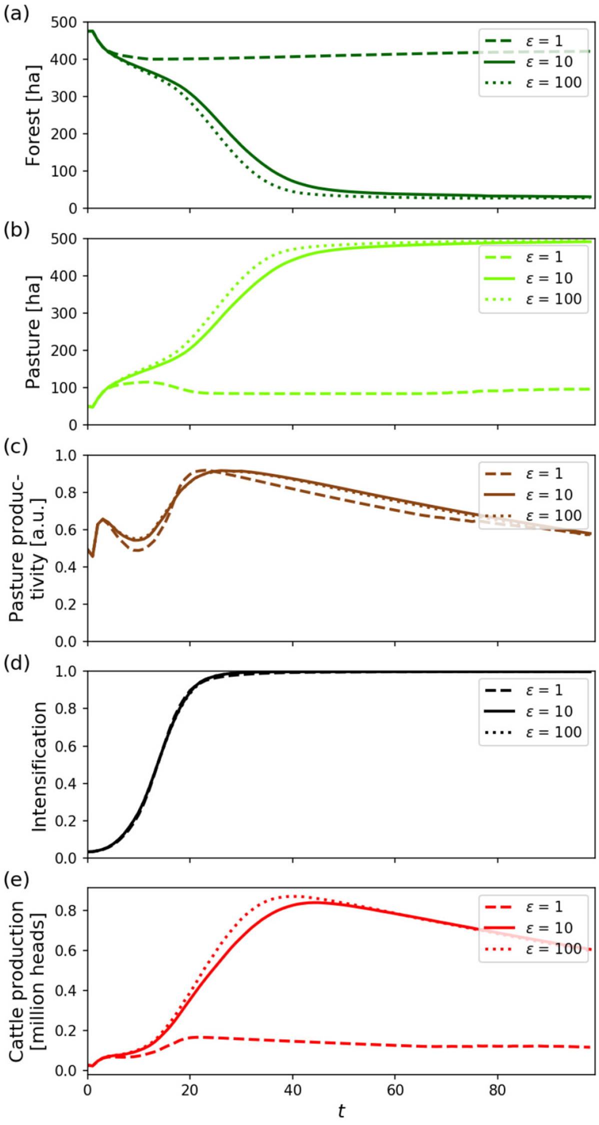


Fig. 9. Sensitivities for variations of the price elasticity of demand ε.

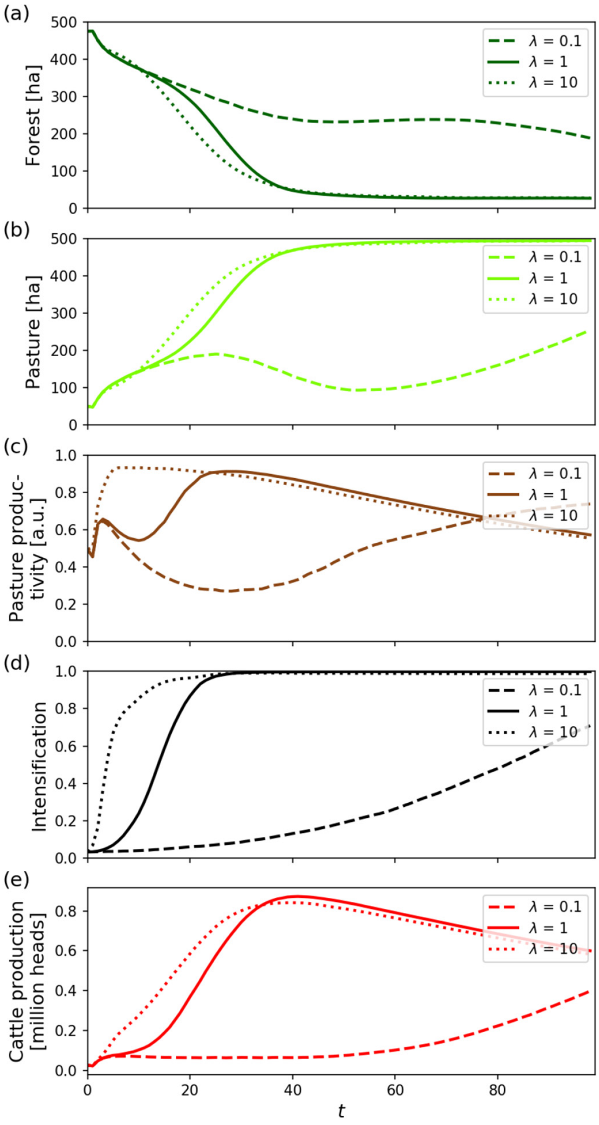


Fig. 10. Sensitivities for variations of the imitation rate λ.

them in a relevant way.

4. Discussion

The model analysis above showed that already a stylized model including a few feedbacks and representing the heterogeneity of agents yields rich non-linear dynamics. The model design implies that only price effects, limited access to credit, high costs for investments, and constraints on decision making impede total deforestation in the *abacra* model. For these assumptions, we find that deforestation can only be curbed by intensification if price elasticity of demand in the model is high and the cattle market saturates at some point.

The elasticity in the model can be interpreted as a measure of integration of the local cattle market into national or international markets. With ongoing globalization and building of infrastructure in the Amazon (de Toledo et al., 2017), the elasticity of demand for local markets rises such that markets will not easily saturate.

Especially with the pavement of the BR-163 highway, our example region around Novo Progresso is increasingly well accessible and connected to the rest of Brazil (Fearnside, 2007). Therefore, a high degree of integration of the local cattle market into national and international markets is probable (Gollnow et al., 2018). In our model, this is represented by a high elasticity of demand approximating a purely price-taking supply side. However, there may be differences also within the region, for example regarding the accessibility of properties far away from the highway (Weinhold and Reis, 2008).

We can similarly interpret the share of teleconnections in the network: with ongoing technical progress, the interaction between ranchers that are not located in the same neighborhood will increase. The model results suggest that this only has a minor effect on the deforestation outcomes. Furthermore, if the costs for intensification are high, limitations on credit hamper the increase of deforestation in the model. This may reflect the success of policies limiting access to agricultural credit in municipalities with high deforestation rates (Assunção et al., 2013).

The model analysis indicates that the exact trajectories depend on

Reproduced from: F. Müller-Hansen et al., Can Intensification of Cattle Ranching Reduce Deforestation in the Amazon? Insights From an Agent-based Social-Ecological Model, Ecol. Econ., vol. 159, pp. 198-211, 2019, doi: 10.1016/j.ecolecon.2018.12.025, Copyright 2019 Elsevier B.V. All rights reserved, with permission from Elsevier.

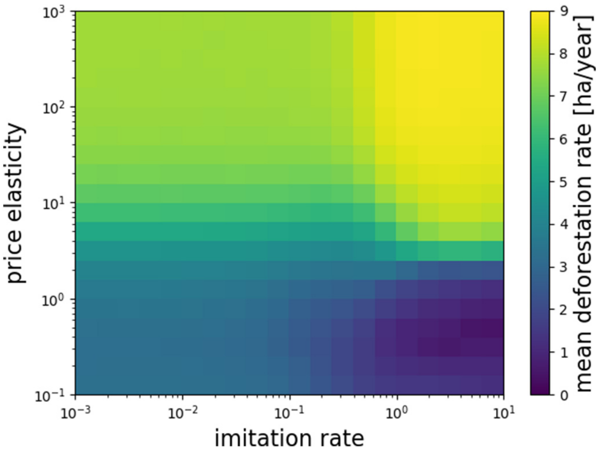


Fig. 11. Average deforestation per year and property in dependence on price elasticity and imitation rate. Parameters are given in Table 1 and the initial conditions are based on deforested areas in the study area by 2000. The displayed values are the ensemble median over 100 runs.

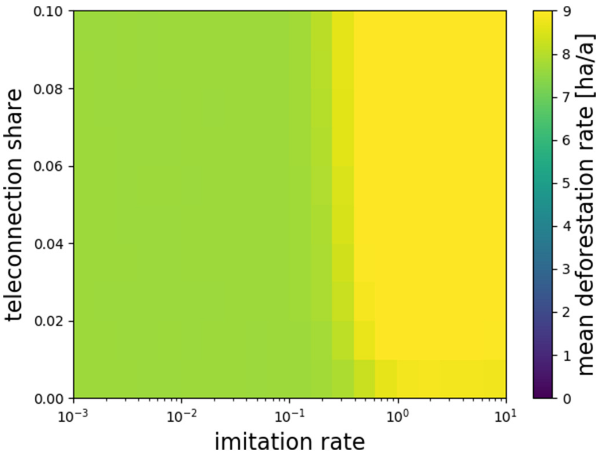


Fig. 12. Average deforestation per year and property in dependence on teleconnection share α and imitation rate λ. Parameters as in Table 1 with $\varepsilon = 100$.

the parameterization of the implemented decision processes and initial conditions. The decision rules used in this model are derived from a survey of the literature and are tuned to reproduce observed land-use patterns in the region. However, there are no empirical studies on the motives, goals and decision procedures of agents, which makes it difficult to construct sound decision functions. Further research in this direction is needed to improve the validity of model results, especially the collection of evidence on how agents in frontier regions make decisions about land use. Furthermore, there often remain many indeterminacies when deriving decision rules from empirical observations even if plenty of data is available. This gap can be bridged by comparing different decision making strategies of agents in a model with empirical data, for instance using inter-temporal or myopic optimization, satisficing, and individual learning approaches. Separating between single intensification practices and techniques would furthermore result in a characterization of intensification as a continuous process, helping to answer for instance the question which level of intensification would be individually and socially optimal.

To date, intensification of cattle ranching in the Amazon is slow (Sparovek et al., 2018). Especially in remote areas, there is limited access to transportation infrastructure, energy, and labor. Furthermore, the land tenure system and land market play an important role for deforestation dynamics because deforestation is a means for agents to lay claim to land and later get land titles through regularization processes (Barretto et al., 2013; Sparovek et al., 2015). This can make deforestation a speculative investment. We did not account for these factors in the *abacra* model, but future extensions focusing on any of these issues could be used to investigate their interplay with intensification further.

The adoption of intensification techniques for cattle production also generates new environmental problems not captured in our model. Intensified systems are associated with heavy nitrogen pollution, water usage, and soil depletion (Tilman et al., 2011). Including such impacts into the model would allow analyzing the environmental trade-offs between intensified and extensive cattle production further. The aim of such modeling could be to identify agricultural practices that are both economically viable and sustainable over long time scales.

In the past, Brazilian conservation policies like the extension of legal reserves from 50 to 80% of private lands in 1996 (Alston and Mueller, 2007) and the monitoring and sanctioning of deforestation activities have reduced deforestation considerably (Nepstad et al., 2014). But current legislation provides low incentives for full compliance with the law, especially regarding reforestation (Azevedo et al., 2017). The internationalization of agricultural commodity markets increased pressure on producers to comply with environmental legislation (Nepstad et al., 2006). This resulted in industry initiatives to monitor compliance such as the zero-deforestation agreement from 2009 (Gibbs et al., 2016). However, recent research shows that the positive effect of the zero-deforestation agreement is undermined by leakage effects ("cattle laundering" Alix-Garcia and Gibbs, 2017; Klingler et al., 2018). To effectively exclude violators of environmental law from the beef supply chain, monitoring of the entire life cycle of cattle would be necessary.

Measures to foster land-use intensification have been debated as an alternative anti-deforestation policy. However, as Merry and Soares-Filho (2017) convincingly argued, intensification policies alone will not lead to better conservation outcomes, i.e., less deforestation. Intensification is rather the result of effective conservation policies. This is consistent with our model results for well integrated markets. Given the model results in this study and despite the limitations of our model, we conclude that anti-deforestation policies only aiming at intensification of cattle ranching will not have the desired result if they are not accompanied by measures that limit the agents' access to new land. Policies aiming to increase intensification cannot replace conservation policies.

An important issue for the design of future anti-deforestation policies is the huge heterogeneity of actors in frontier development. The roles of various types of agents with respect to deforestation outcomes changes as a response to new policy implementations and their effectiveness. Recent studies comparing the contributions of small-holders and large land-owners found opposing trends, depending on the time and location they focused on (Godar et al., 2014, 2012; Richards and VanWey, 2015). For example, large-scale ranchers, who drive land concentration in more consolidated areas, are susceptible to other incentives than small-holders in remote areas, mainly involved in subsistence farming. To investigate the different effect of intensification policies and economic drivers on this heterogeneity of agents is a challenge for future modeling studies.

In general, development and environmental policies for the Amazon have to face the various trade-offs between social and environmental issues (de Toledo et al., 2017). Cattle ranching remains an important source of income for land holders in the Amazon. As the demand for cattle products is increasing world-wide (Thornton, 2010), ranching provides an economic perspective for the region. Policies have to guarantee that local incomes are maintained or increased while conserving the ecosystems. Therefore, it is essential that they can anticipate the multiple feedbacks in the system that could undermine the effectiveness of policies. It remains an open question how cattle ranching in

Reproduced from: F. Müller-Hansen et al., Can Intensification of Cattle Ranching Reduce Deforestation in the Amazon? Insights From an Agent-based Social-Ecological Model, Ecol. Econ., vol. 159, pp. 198-211, 2019, doi: 10.1016/j.ecolecon.2018.12.025, Copyright 2019 Elsevier B.V. All rights reserved, with permission from Elsevier.

the Amazon will become an environmentally and socially sustainable economic activity in the long term, with or without intensification.

5. Conclusion

This study presents and analyzes a new agent-based model that conceptualizes the intensification of cattle ranching as a socially mediated process. With this approach, we shed light on the interplay between ecological dynamics, economic conditions, decision making of agents, and interactions on a social network. We show how even from very stylized assumptions about these dynamics, a rich non-linear behavior arises at the system level, which can be explained by the various feedback loops between them. We use recent data sets on land properties (CAR) and deforestation (PRODES) in a frontier region to demonstrate the model dynamics for specific initializations and parameterizations.

In particular, we highlight the effect of the imitation rate and price elasticity of demand for cattle. We show that higher imitation rates, which lead to faster intensification, can only reduce deforestation in a market that saturates. On the other hand, under conditions of less responsive prices, faster intensification can even lead to higher deforestation. Our model shows these effects on a regional scale but similar rebound effects have been discussed for the global food system (Lambin and Meyfroidt, 2011).

The model presented here is only a first step towards including local social interaction into models of land-use change in the context of tropical deforestation. Future work with agent-based models could focus on evaluating the effectiveness and resilience of anti-deforestation policies accounting for heterogeneities of actors in the deforestation process (Godar et al., 2014). Agent-based models are a powerful tool for such analyses because they can represent heterogeneities and account for the various feedbacks in the system. Thereby, they might help developing an economic perspective for the region that provides improvements in livelihoods and at the same time reduce deforestation.

Acknowledgments

This work was developed within the scope of the IRTG 1740/TRP 2011/50151-0 and 2015/50122-0 funded by the Deutsche Forschungsgemeinschaft (DFG), Germany and the São Paulo Research Foundation FAPESP, Brazil, whose financial support is thankfully acknowledged. J.F.D. is grateful for financial support by the Stordalen Foundation (via the Planetary Boundary Research Network PB.net), the EarthLeague's EarthDoc program, and the Leibniz Association (project DOMINOES). F.M.-H. thanks Tim Kittel, Catrin Ciemer and Silvana Tiedemann, the members of the ECOSTAB and copan flagships at the Potsdam Institute for Climate Impact Research (PIK), and researchers of CCST at INPE for fruitful discussions. Special thanks go to the three anonymous reviewers whose comments led to substantial improvements of the manuscript. Last but not least, we want to thank all contributors and developers of the python ecosystem that we used to implement the model. Parameter studies were performed on the high-performance computer system of PIK, supported by the European Regional Development Fund, BMBF, and the Land Brandenburg.

Declaration of interest. The authors declare that they have no conflict of interest.

Code and data availability. The model code is available from github.com/fmhansen/abacra. The data used for this study is available online and referenced accordingly.

Appendix A

Table 2
Overview of variables, symbols, and units in the model.

Variable	Symbol	Unit
Pasture area	P_t	ha
Forest area	F_t	ha
Secondary vegetation area	S_t	ha
Pasture productivity	q_t	a.u.
Secondary vegetation productivity	v_t	a.u.
Savings of rancher	k_t	BRL
Income	I_t	BRL
Consumption	C_t	BRL
Deforestation	d_t	ha/year
Abandonment	a_t	ha/year
Reuse	r_t	ha/year
Management effort	m_t	a.u.
Stocking rate for pasture	l_t	Head/ha

Appendix B. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.ecolecon.2018.12.025.

References

Aguiar, A.P.D., Ometto, J.P., Nobre, C., Lapola, D.M., Almeida, C., Vieira, I.C., Soares, J.V., Alvala, R., Saatchi, S., Valeriano, D., Castilla-Rubio, J.C., 2012. Modeling the spatial and temporal heterogeneity of deforestation-driven carbon emissions: the INPE-EM framework applied to the Brazilian Amazon. Glob. Chang. Biol. 18, 3346–3366. https://doi.org/10.1111/j.1365-2486.2012.02782.x.

Aguiar, A.P.D., Vieira, I.C.G., Assis, T.O., Dalla-Nora, E.L., Toledo, P.M., Santos-Junior, R.A.O., Batistella, M., Coelho, A.S., Savaget, E.K., Aragão, L.E.O.C., Nobre, C.A., Ometto, J.P.H., 2016. Land use change emission scenarios: anticipating a forest transition process in the Brazilian Amazon? Glob. Chang. Biol. 22, 1821–1840. https://doi.org/10.1111/gcb.13134.

Alix-Garcia, J., Gibbs, H.K., 2017. Forest conservation effects of Brazil's zero deforestation cattle agreements undermined by leakage. Glob. Environ. Chang. 47, 201–217. https://doi.org/10.1016/j.gloenvcha.2017.08.009.

Almeida, C.A., Coutinho, A.C., Esquerdo, J.C.D.M., Adami, M., Venturieri, A., Diniz, C.G., Dessay, N., Durieux, L., Gomes, A.R., 2016. High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data. Acta Amazon. 46, 291–302. https://doi.org/10.1590/1809-4392201505504.

Alston, L.J., Mueller, B., 2007. Legal reserve requirements in Brazilian forests: path dependent evolution of de facto legislation. Rev. Econ. 8, 25–53.

An, L., 2012. Modeling human decisions in coupled human and natural systems: review of agent-based models. Ecol. Model. 229, 25–36. https://doi.org/10.1016/j.ecolmodel.2011.07.010.

Andersen, L.E., Groom, B., Killick, E., Ledezma, J.C., Palmer, C., Weinhold, D., 2017. Modelling land use, deforestation, and policy: a hybrid optimisation-heterogeneous agent model with application to the Bolivian Amazon. Ecol. Econ. 135, 76–90. https://doi.org/10.1016/j.ecolecon.2016.12.033.

Angelsen, A., Kaimowitz, D., 1999. Rethinking the causes of deforestation: lessons from economic models. World Bank Res. Obs. 14, 73–98. https://doi.org/10.1093/wbro/14.1.73.

Angelsen, A., Kaimowitz, D., 2001. Introduction: the role of agricultural technologies in tropical deforestation. In: Angelsen, A., Kaimowitz, D. (Eds.), Agricultural Technologies and Tropical Deforestation. CABI Publishing, Oxon, UK, and New York, pp. 1–17.

Assunção, J., Gandour, C., Rocha, R., Rocha, R., 2013. Does credit affect deforestation? Evidence from a rural credit policy in the Brazilian Amazon. In: Technical Report, Climate Policy Initiative. https://climatepolicyinitiative.org/publication/does-credit-affect-deforestation-evidence-from-a-rural-credit-policy-in-the-brazilian-amazon/.

Azevedo, A.A., Rajão, R., Costa, M.A., Stabile, M.C.C., Macedo, M.N., dos Reis, T.N.P., Alencar, A., Soares-Filho, B.S., Pacheco, R., 2017. Limits of Brazil's forest code as a means to end illegal deforestation. Proc. Natl. Acad. Sci. U. S. A. 114, 7653–7658. https://doi.org/10.1073/pnas.1604768114.

Barona, E., Ramankutty, N., Hyman, G., Coomes, O.T., 2010. The role of pasture and soybean in deforestation of the Brazilian Amazon. Environ. Res. Lett. 5, 024002. https://doi.org/10.1088/1748-9326/5/2/024002.

Barretto, A.G., Berndes, G., Sparovek, G., Wirsenius, S., 2013. Agricultural intensification in Brazil and its effects on land-use patterns: an analysis of the 1975–2006 period. Glob. Chang. Biol. 19, 1804–1815. https://doi.org/10.1111/gcb.12174.

Berger, T., 2001. Agent-based spatial models applied to agriculture: a simulation tool for technology diffusion, resource use changes and policy analysis. Agric. Econ. 25, 245–260. https://doi.org/10.1111/j.1574-0862.2001.tb00205.x.

Bowman, M.S., Soares-Filho, B.S., Merry, F.D., Nepstad, D.C., Rodrigues, H., Almeida, O.T., 2012. Persistence of cattle ranching in the Brazilian Amazon: a spatial analysis of the rationale for beef production. Land Use Policy 29, 558–568. https://doi.org/10.1016/j.landusepol.2011.09.009.

Busch, J., Ferretti-Gallon, K., 2017. What drives deforestation and what stops it? A meta-analysis. Rev. Environ. Econ. Policy 11, 3–23. https://doi.org/10.1093/reep/rew013.

Cano-Crespo, A., Oliveira, P.J., Boit, A., Cardoso, M., Thonicke, K., 2015. Forest edge burning in the Brazilian Amazon promoted by escaping fires from managed pastures. J. Geophys. Res. Biogeosci. 120, 2095–2107. https://doi.org/10.1002/2015JG002914.

CAR Sistema Nacional de Cadastro Ambiental Rural - Base de Downloads. http://www.car.gov.br/publico/municipios/downloads, Accessed date: 2 November 2018.

Cohn, A., Bowman, M., Zilberman, D., 2011. The viability of cattle ranching intensification in Brazil as a strategy to spare land and mitigate greenhouse gas emissions. CCAFS Working Paper No. 11. http://hdl.handle.net/10568/10722.

Cohn, A.S., Mosnier, A., Havlík, P., Valin, H., Herrero, M., Schmid, E., O'Hare, M., Obersteiner, M., 2014. Cattle ranching intensification in Brazil can reduce global greenhouse gas emissions by sparing land from deforestation. Proc. Natl. Acad. Sci. U. S. A. 111, 7236–7241. https://doi.org/10.1073/pnas.1307163111.

Costa, S.S., 2012. Regional Scale Agent-Based Modelling of Land Change: Evolving Institutional Arrangements in Frontier Areas. Ph.D. thesis. INPE, São José dos Campos.

Currarini, S., Marchiori, C., Tavoni, A., 2016. Network economics and the environment: insights and perspectives. Environ. Resour. Econ. 65, 159–189. https://doi.org/10.1007/s10640-015-9953-6.

Davidson, E.A., De Carvalho, C.J., Figueira, A.M., Ishida, F.Y., Ometto, J.P.H., Nardoto, G.B., Sabá, R.T., Hayashi, S.N., Leal, E.C., Vieira, I.C.G., Martinelli, L.A., 2007. Recuperation of nitrogen cycling in Amazonian forests following agricultural abandonment. Nature 447, 995–998. https://doi.org/10.1038/nature05900.

de Toledo, P.M., Dalla-Nora, E., Vieira, I.C.G., Aguiar, A.P.D., Araújo, R., 2017. Development paradigms contributing to the transformation of the Brazilian Amazon: do people matter? Curr. Opin. Environ. Sustain. 26–27, 77–83. https://doi.org/10.1016/j.cosust.2017.01.009.

Deadman, P., Robinson, D., Moran, E., Brondizio, E., 2004. Colonist household decisionmaking and land-use change in the Amazon rainforest: an agent-based simulation. Environ. Plann. B. Plann. Des. 31, 693–709. https://doi.org/10.1068/b3098.

Erb, K.-H., Fetzel, T., Kastner, T., Kroisleitner, C., Lauk, C., Mayer, A., Niedertscheider, M., 2016. Livestock grazing, the neglected land use. In: Haberl, H., Fischer-Kowalski, M., Krausmann, F., Winiwarter, V. (Eds.), Social Ecology. Human-Environment Interactions. 5. Springe, Cham, pp. 295–310. https://doi.org/10.1007/978-3-319-33326-7_13.

Fearnside, P.M., 2007. Brazil's Cuiabá-Santarém (BR-163) highway: the environmental cost of paving a soybean corridor through the Amazon. Environ. Manag. 39, 601–614. https://doi.org/10.1007/s00267-006-0149-2.

Feder, G., Umali, D.L., 1993. The adoption of agricultural innovations. A review. Technol. Forecast. Soc. Chang. 43, 215–239. https://doi.org/10.1016/0040-1625(93)90053-A.

FGVIBRE FGVDados. URL: http://portalibre.fgv.br/main.jsp?lumChannelId=402880811D8E34B9011D92C493F131B2, (accessed: November 2, 2018).

Garrett, R.D., Gardner, T.A., Morello, T.F., Marchand, S., Barlow, J., de Blas, D.E., Ferreira, J., Lees, A.C., Parry, L., 2017. Explaining the persistence of low income and environmentally degrading land uses in the Brazilian Amazon explaining the persistence of low income and environmentally degrading land uses in the Brazilian Amazon. Ecol. Soc. 22, 27. https://doi.org/10.5751/ES-09364-220327.

Gibbs, H.K., Munger, J., L'Roe, J., Barreto, P., Pereira, R., Christie, M., Amaral, T., Walker, N.F., 2016. Did ranchers and slaughterhouses respond to zero-deforestation agreements in the Brazilian Amazon? Conserv. Lett. 9, 32–42. https://doi.org/10.1111/conl.12175.

Gigerenzer, G., Gaissmaier, W., 2011. Heuristic decision making. Annu. Rev. Psychol. 62, 451–482. https://doi.org/10.1146/annurev-psych-120709-145346.

Godar, J., Gardner, T.A., Tizado, E.J., Pacheco, P., 2014. Actor-specific contributions to the deforestation slowdown in the Brazilian Amazon. Proc. Natl. Acad. Sci. U. S. A. 111, 15591–15596. https://doi.org/10.1073/pnas.1322825111.

Godar, J., Tizado, E.J., Pokorny, B., 2012. Who is responsible for deforestation in the Amazon? A spatially explicit analysis along the Transamazon Highway in Brazil. For. Ecol. Manag. 267, 58–73. https://doi.org/10.1016/j.foreco.2011.11.046.

Gollnow, F., Göpel, J., deBarros Viana Hissa, L., Schaldach, R., Lakes, T., 2018. Scenarios of land-use change in a deforestation corridor in the Brazilian Amazon: combining two scales of analysis. Reg. Environ. Chang. 18, 143–159. https://doi.org/10.1007/s10113-017-1129-1.

Groeneveld, J., Müller, B., Buchmann, C., Dressler, G., Guo, C., Hase, N., Hoffmann, F., John, F., Klassert, C., Lauf, T., Liebelt, V., Nolzen, H., Pannicke, N., Schulze, J., Weise, H., Schwarz, N., 2017. Theoretical foundations of human decision-making in agent-based land use models. A review. Environ. Model Softw. 87, 39–48. https://doi.org/10.1016/j.envsoft.2016.10.008.

Heppenstall, A.J., Heppenstall, A.T., See, L.M., Batty, M. (Eds.), 2012. Agent-Based Models of Geographical Systems. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-8927-4.

Hoelle, J., 2011. Convergence on cattle: political ecology, social group perceptions, and socioeconomic relationships in Acre, Brazil. Culture, Agriculture, Food and Environment 33, 95–106. https://doi.org/10.1111/j.2153-9561.2011.01053.x.

IMEA Custo de Produção da Bovinocultura de Corte. http://www.imea.com.br/imea-site/relatorios-mercado, (accessed: November 2, 2018).

Kaimowitz, D., Angelsen, A., 2008. Will livestock intensification help save Latin America's tropical forests? J. Sustain. For. 27, 6–24. https://doi.org/10.1080/10549810802225168.

Klingler, M., Richards, P.D., Ossner, R., 2018. Cattle vaccination records question the impact of recent zero-deforestation agreements in the Amazon. Reg. Environ. Chang. 18, 33–46. https://doi.org/10.1007/s10113-017-1234-1.

Lambin, E.F., Meyfroidt, P., 2011. Global land use change, economic globalization, and the looming land scarcity. Proc. Natl. Acad. Sci. U. S. A. 108, 3465–3472. https://doi.org/10.1073/pnas.1100480108.

Landers, J.N., 2007. Tropical crop-livestock systems in conservation agriculture: the Brazilian experience. In: Integrated Crop Management Vol. 5-2007. Food and Agriculture Organization of the United Nations, Rome. http://www.fao.org/3/a-a1083e.pdf.

Latawiec, A.E., Strassburg, B.B., Valentim, J.F., Ramos, F., Alves-Pinto, H., 2014. Intensification of cattle ranching production systems: socioeconomic and environmental synergies and risks in Brazil. Animal 8, 1255–1263. https://doi.org/10.1017/S1751731114001566.

Lenton, T.M., Held, H., Kriegler, E., Hall, J.W., Lucht, W., Rahmstorf, S., Schellnhuber, H.J., 2008. Tipping elements in the Earth's climate system. Proc. Natl. Acad. Sci. U. S. A. 105, 1786–1793. https://doi.org/10.1073/pnas.0705414105.

Maertens, A., Barrett, C.B., 2012. Measuring social networks' effects on agricultural technology adoption. Am. J. Agric. Econ. 95, 353–359. https://doi.org/10.1093/ajae/aas049.

Manson, S.M., Evans, T., 2007. Agent-based modeling of deforestation in southern Yucatan, Mexico, and reforestation in the Midwest United States. Proc. Natl. Acad. Sci. U. S. A. 104 (52), 20678–20683. https://doi.org/10.1073/pnas.0705802104.

Matthews, R.B., Gilbert, N.G., Roach, A., Polhill, J.G., Gotts, N.M., 2007. Agent-based land-use models: a review of applications. Landsc. Ecol. 22, 1447–1459. https://doi.org/10.1007/s10980-007-9135-1.

Mena, C.F., Walsh, S.J., Frizzelle, B.G., Xiaozheng, Y., Malanson, G.P., 2011. Land use change on household farms in the Ecuadorian Amazon: design and implementation of an agent-based model. Appl. Geogr. 31, 210–222. https://doi.org/10.1016/j.apgeog.2010.04.005.

Merry, F., Soares-Filho, B., 2017. Will intensification of beef production deliver conservation outcomes in the Brazilian Amazon? Elementa 5, 24. https://doi.org/10.1525/elementa.224.

Michetti, M., 2012. Modelling land use, land-use change, and forestry in climate change: a review of major approaches. FEEM Working Paper No. 46.2012. https://doi.org/10.2139/ssrn.2122298.

Müller, B., Bohn, F., Dreßler, G., Groeneveld, J., Klassert, C., Martin, R., Schlüter, M., Schulze, J., Weise, H., Schwarz, N., 2013. Describing human decisions in agent based models - ODD + D, an extension of the ODD protocol. Environ. Model Softw. 48, 37–48. https://doi.org/10.1016/j.envsoft.2013.06.003.

Müller-Hansen, F., Schlüter, M., Mäs, M., Donges, J.F., Kolb, J.J., Thonicke, K., Heitzig, J., 2017. Towards representing human behavior and decision making in Earth system models an overview of techniques and approaches. Earth Syst. Dynam. 8, 977–1007. https://doi.org/10.5194/esd-8-977-2017.

Myers, R.J.K., Robbins, G.B., 1991. Sustaining productive pasture in the tropics - 5. Maintaining productive sown grass pastures. Tropical Grasslands 25, 104–110.

Nepstad, D., McGrath, D., Stickler, C., Alencar, A., Azevedo, A., Swette, B., Bezerra, T., DiGiano, M., Shimada, J., Seroa da Motta, R., Armijo, E., Castello, L., Brando, P., Hansen, M.C., McGrath-Horn, M., Carvalho, O., Hess, L., 2014. Slowing Amazon deforestation through public policy and interventions in beef and soy supply chains. Science 344, 1118–1123. https://doi.org/10.1126/science.1248525.

Nepstad, D.C., Stickler, C.M., Almeida, O.T., 2006. Globalization of the Amazon soy and beef industries: opportunities for conservation. Conserv. Biol. 20, 1595–1603. https://doi.org/10.1111/j.1523-1739.2006.00510.x.

Oliveira, L.J.C., Costa, M.H., Soares-Filho, B.S., Coe, M.T., 2013. Large-scale expansion of agriculture in Amazonia may be a no-win scenario. Environ. Res. Lett. 8, 024021. https://doi.org/10.1088/1748-9326/8/2/024021.

Pacheco, P., 2012. Actor and frontier types in the Brazilian Amazon: assessing interactions and outcomes associated with frontier expansion. Geoforum 43, 864–874.

https://doi.org/10.1016/j.geoforum.2012.02.003.

Pacheco, P., Poccard-Chapuis, R., 2012. The complex evolution of cattle ranching development amid market integration and policy shifts in the Brazilian Amazon. Ann. Assoc. Am. Geogr. 102, 1366–1390. https://doi.org/10.1080/00045608.2012.678040.

Parker, D.C., Entwisle, B., Rindfuss, R.R., Vanwey, L.K., Manson, S.M., Moran, E., An, L., Deadman, P., Evans, T.P., Linderman, M., Mussavi Rizi, S.M., Malanson, G., 2008. Case studies, cross-site comparisons, and the challenge of generalization: comparing agent-based models of land-use change in frontier regions. J. Land Use Sci. 3, 41–72. https://doi.org/10.1080/17474230802048151.

Parker, D.C., Manson, S.M., Janssen, M.A., Hoffmann, M.J., Deadman, P., 2003. Multi-agent systems for the simulation of land-use and land-cover change: a review. Ann. Assoc. Am. Geogr. 93, 314–337. https://doi.org/10.1111/1467-8306.9302004.

Perz, S., Skole, D., 2003a. Secondary forest expansion in the Brazilian Amazon and the refinement of forest transition theory. Soc. Nat. Resour. 16, 277–294. https://doi.org/10.1080/08941920390178856.

Perz, S., Skole, D.L., 2003b. Social determinants of secondary forests in the Brazilian Amazon. Soc. Sci. Res. 32, 25–60. https://doi.org/10.1016/s0049-089x(02)00012-1.

Poorter, L., Bongers, F., Aide, T.M., Almeyda Zambrano, A.M., Balvanera, P., Becknell, J.M., Boukili, V., Brancalion, P.H., Broadbent, E.N., Chazdon, R.L., Craven, D., De Almeida-Cortez, J.S., Cabral, G.A., De Jong, B.H., Denslow, J.S., Dent, D.H., DeWalt, S.J., Dupuy, J.M., Durán, S.M., Espírito-Santo, M.M., Fandino, M.C., César, R.G., Hall, J.S., Hernandez-Stefanoni, J.L., Jakovac, C.C., Junqueira, A.B., Kennard, D., Letcher, S.G., Licona, J.C., Lohbeck, M., Marín-Spiotta, E., Martínez-Ramos, M., Massoca, P., Meave, J.A., Mesquita, R., Mora, F., Munõz, R., Muscarella, R., Nunes, Y.R., Ochoa-Gaona, S., De Oliveira, A.A., Orihuela-Belmonte, E., Penã-Claros, M., Pérez-Garciá, E.A., Piotto, D., Powers, J.S., Rodríguez-Vel'azquez, J., Romero-Pérez, I.E., Ruíz, J., Saldarriaga, J.G., Sanchez-Azofeifa, A., Schwartz, N.B., Steininger, M.K., Swenson, N.G., Toledo, M., Uriarte, M., Van Breugel, M., Van Der Wal, H., Veloso, M.D., Vester, H.F., Vicentini, A., Vieira, I.C., Bentos, T.V., Williamson, G.B., Rozendaal, D.M., 2016. Biomass resilience of Neotropical secondary forests. Nature 530, 211–214. https://doi.org/10.1038/nature16512.

PRODES Projeto de Monitoramento da Floresta Amazônica Brasileira por Satélite. URL: http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes, (accessed: November 2, 2018).

Quaas, M.F., Baumgärtner, S., Becker, C., Frank, K., Müller, B., 2007. Uncertainty and sustainability in the management of rangelands. Ecol. Econ. 62, 251–266. https://doi.org/10.1016/j.ecolecon.2006.03.028.

Richards, P., Arima, E., Vanwey, L., Cohn, A., Bhattarai, N., 2017. Are Brazil's deforesters avoiding detection? Conserv. Lett. 10, 470–476. https://doi.org/10.1111/conl.12310.

Richards, P.D., VanWey, L., 2015. Farm-scale distribution of deforestation and remaining forest cover in Mato Grosso. Nat. Clim. Chang. 6, 418–425. https://doi.org/10.1038/nclimate2854.

Richards, P.D., Walker, R.T., Arima, E.Y., 2014. Spatially complex land change: the indirect effect of Brazil's agricultural sector on land use in Amazonia. Glob. Environ. Chang. 29, 1–9. https://doi.org/10.1016/j.gloenvcha.2014.06.011.

Robalino, J.A., Pfaff, A., 2012. Contagious development: neighbor interactions in deforestation. J. Dev. Econ. 97, 427–436. https://doi.org/10.1016/j.jdeveco.2011.06.003.

Salisbury, D.S., Schmink, M., 2007. Cows versus rubber: changing livelihoods among Amazonian extractivists. Geoforum 38, 1233–1249. https://doi.org/10.1016/j.geoforum.2007.03.005.

Satake, A., Rudel, T.K., 2007. Modeling the forest transition: forest scarcity and ecosystem service hypotheses. Ecol. Appl. 17, 2024–2036. https://doi.org/10.1890/07-0283.1.

Schlüter, M., Mcallister, R.R.J., Arlinghaus, R., Bunnefeld, N., Eisenack, K., Hölker, F., Milner-Gulland, E.J., Müller, B., Nicholson, E., Quaas, M., Stöven, M., 2012. New horizons for managing the environment: a review of coupled social-ecological systems modeling. Nat. Resour. Model. 25, 219–272. https://doi.org/10.1111/j.1939-7445.2011.00108.x.

SEAB Preços médios nominais mensais recibidos polos produtores - boi gordo. URL: http://www.agricultura.pr.gov.br/modules/conteudo/conteudo.php?conteudo=195, (accessed: November 2, 2018).

Serrão, E.A.S., Toledo, J.M., Falesi, I.C., de Veiga, J.B., Teixeira Neto, J.F., 1979. Productivity of cultivated pasture on the the low-fertility soils in the Amazon of Brazil. In: Pasture Production in Acid Soils of the Tropics. CIAT, pp. 195–226.

Soares-Filho, B., Rajão, R., Macedo, M., Carneiro, A., Costa, W., Coe, M., Rodrigues, H., Alencar, A., 2014. Cracking Brazil's forest code. Science 344, 363–364. https://doi.org/10.1126/science.1246663.

Soler, L.S., Verburg, P.H., Alves, D.S., 2014. Evolution of land use in the Brazilian Amazon: from frontier expansion to market chain dynamics. Land 3, 981–1014. https://doi.org/10.3390/land3030981.

Sparovek, G., Barretto, A.G.D.O.P., Matsumoto, M., Berndes, G., 2015. Effects of governance on availability of land for agriculture and conservation in Brazil. Environ. Sci. Technol. 49, 10285–10293. https://doi.org/10.1021/acs.est.5b01300.

Sparovek, G., Guidotti, V., Pinto, L.F.G., Berndes, G., Barretto, A., Cerignoni, F., 2018. Asymmetries of cattle and crop productivity and efficiency during Brazil's agricultural expansion from 1975 to 2006. Elem. Sci. Anth. 6, 25. https://doi.org/10.1525/elementa.187.

The World Bank, Gross Domestic Savings. URL: https://data.worldbank.org/indicator/NY.GDS.TOTL.ZS, (accessed: November 2, 2018).

Thornton, P.K., 2010. Livestock production: recent trends, future prospects. Philos. Trans. R. Soc., B 365, 2853–2867. https://doi.org/10.1098/rstb.2010.0134.

Tilman, D., Balzer, C., Hill, J., Befort, B.L., 2011. Global food demand and the sustainable intensification of agriculture. Proc. Natl. Acad. Sci. 108, 20260–20264. https://doi.org/10.1073/pnas.1116437108.

Traulsen, A., Semmann, D., Sommerfeld, R.D., Krambeck, H.-J., Milinski, M., 2010. Human strategy updating in evolutionary games. Proc. Natl. Acad. Sci. U. S. A. 107, 2962–2966. https://doi.org/10.1073/pnas.0912515107.

Van Kampen, N.G., 2007. Stochastic Processes in Physics and Chemistry, 3rd ed. North Holland, Amsterdam.

Verburg, P.H., Soepboer, W., Veldkamp, A., Limpiada, R., Espaldon, V., Mastura, S.S.A., 2002. Modeling the spatial dynamics of regional land use: the CLUE-S model. Environ. Manag. 30, 391–405. https://doi.org/10.1007/s00267-002-2630-x.

Waxman, B.M., 1988. Routing of multipoint connections. IEEE J. Sel. Areas Commun. 6, 1617–1622. https://doi.org/10.1109/49.12889.

Weinhold, D., Reis, E., 2008. Transportation costs and the spatial distribution of land use in the Brazilian Amazon. Glob. Environ. Chang. 18, 54–68. https://doi.org/10.1016/j.gloenvcha.2007.06.004.

West, T.A., Grogan, K.A., Swisher, M.E., Caviglia-Harris, J.L., Sills, E., Harris, D., Roberts, D., Putz, F.E., 2018. A hybrid optimization-agent-based model of REDD+ payments to households on an old deforestation frontier in the Brazilian Amazon. Environ. Model. Softw. 100, 159–174. https://doi.org/10.1016/j.envsoft.2017.11.007.

Wiedermann, M., Donges, J.F., Heitzig, J., Lucht, W., Kurths, J., 2015. Macroscopic description of complex adaptive networks coevolving with dynamic node states. Phys. Rev. E 91, 052801. https://doi.org/10.1103/PhysRevE.91.052801.

Zemp, D.C., Schleussner, C.F., Barbosa, H.M., Rammig, A., 2017. Deforestation effects on Amazon forest resilience. Geophys. Res. Lett. 44, 6182–6190. https://doi.org/10.1002/2017GL072955.

zu Ermgassen, E.K., de Alcântara, M.P., Balmford, A., Barioni, L., Neto, F.B., Bettarello, M.M., de Brito, G., Carrero, G.C., Florence, E.d.A., Garcia, E., Gonçalves, E.T., da Luz, C.T., Mallman, G.M., Strassburg, B.B., Valentim, J.F., Latawiec, A., 2018. Results from on-the-ground efforts to promote sustainable cattle ranching in the Brazilian Amazon. Sustainability 10, 1301. https://doi.org/10.3390/su10041301.

Copan members

Philipp S. Arndt was an undergraduate summer intern at *copan* in 2015, analyzing the 'Great Acceleration' time series with a focus on the Anthropocene discussion. He is now a doctoral student at Scripps Institution of Oceanography in San Diego and a NASA "Future Investigator" studying meltwater systems on Antarctic ice shelves using satellite remote sensing techniques. [Arndt et al., 2016]

Yuki Asano did his BSc thesis at *copan* in 2017, building a macro economic agent-based model that includes social dynamics. He is now a PhD student focussing on Computer Vision at the University of Oxford. [Asano et al., 2019]

Wolfram Barfuss did his doctoral dissertation in *copan* from 2015 to 2019, working on multi-agent learning in social-ecological system models. [Barfuss et al., 2020, Barfuss et al., 2019, Barfuss et al., 2018, Barfuss et al., 2017]

Boyan Beronov collaborated with *copan* in 2014–2015, via his BSc thesis in physics about causal entropy in conceptual models of social dynamics, and as a contributor to the Pyunicorn software. He continued with an MSc in computational science and is currently pursuing a PhD in artificial intelligence at the University of British Columbia, Vancouver. [Donges, J. F. and Heitzig, J. et al., 2020]

Jonathan F. Donges is a theoretical physicist and one of the speakers of *copan* since 2013. He also leads PIK's FutureLab on Earth Resilience in the Anthropocene and the working group on Whole Earth System Analysis.

Birte Ewers wrote her master's thesis for her Economics MSc in the *copan* group. Her topic focused on divestment from fossil fuels and agent-based modeling. She is currently working at the Institute for Energy and Environmental Research (ifeu) in Heidelberg. [Ewers, B. and Donges, J. F. et al., 2019]

Fabian Geier did his physics master thesis with *copan* working on opinion formation on multi-layer complex networks in 2016. He went on to become a data scientist and software developer at the consulting company Ramboll. [Geier et al., 2019]

Luzie Helfmann is a PhD student at FU Berlin and in the *copan* group. She is developing methods for the analysis of tipping dynamics in high-dimensional agent-based models. [Helfmann et al., 2021, Helfmann et al., 2020]

Jobst Heitzig is a mathematician and one of the speakers of *copan* since 2013. He also leads PIK's FutureLab on Game Theory and Networks of Interacting Agents.

Tanja Holstein did her physics master's thesis in *copan* in 2019, working on a model of diffusively coupled socio-ecological resource exploitation networks. She is now a PhD student in virus bioinformatics at the Bundesanstalt für Materialforschung und -Prüfung and Ghent University. [Holstein et al., 2021]

Johannes Kassel finished his physics master's thesis in *copan* in 2019, studying a network model for human behavioural changes. Afterwards he started a PhD under the supervision of Holger Kantz at MPI PKS in Dresden. [Donges, J. F. and Heitzig, J. et al., 2020]

Tim Kittel did his PhD on the analysis of socio-ecological models at the *copan* group at PIK. After an intermezzo as entrepreneur he is now a project manager at one of the leading North-European software consultancies. [Kittel et al., 2017a, Kittel et al., 2017b]

Niklas Kitzmann is a PhD student in the *copan* group since 2019. His research focuses on exploring network-based approaches to social tipping points, for example in the spread of sustainability innovations between cities. [Donges, J. F. and Lochner, J. et al., 2021]

Ann Kristin Klose did her environmental modelling bachelor's as well as master's thesis in *copan* in 2017 and 2019–2020, exploring the dynamics of interacting tipping elements in the climate system. [Klose et al., 2020]

Rebekka Koch did an internship at *copan* in 2016 during which she worked on automizing the Topology of Sustainable Management framework. Since 2019, she is a PhD student in theoretical condensed matter physics at the University of Amsterdam. [Kittel et al., 2017b]

Jakob Kolb wrote his dissertation in theoretical physics as part of *copan* and defended it successfully in 2020, studying agent-based modelling and analytic approximations of heuristic decision-making in socio-economic systems. [Kolb et al., 2020, Donges, J.

F. and Heitzig, J. et al., 2020, Asano et al., 2019, Müller-Hansen et al., 2017b]

Till Kolster wrote his master's thesis on the development of an agent-based migration model in the *copan* group in 2017/18. He is now doing his PhD at Siemens and TU Darmstadt on the topic of increasing the german transmission grid's efficiency to integrate more renewables into our energy system. [Donges, J. F. and Heitzig, J. et al., 2020]

Jonathan Krönke did his physics master's thesis in *copan* in 2019 working on modeling and simulation of tipping cascades on complex networks. Since then, he started working as a software developer. [Wunderling et al., 2021, Krönke et al., 2020, Wunderling et al., 2020d]

Jakob Lochner did his physics master's thesis in *copan* in 2019 in which he investigated social contagion dynamics based on empirical network data and stochastic process modeling. [Donges, J. F. and Lochner, J. et al., 2021]

Wolfgang Lucht is co-head of PIK's Research Department on Earth System Analysis and the main initiator and founder of the *copan* idea.

Finn Müller-Hansen wrote his dissertation in theoretical physics as part of *copan* , analyzing deforestation in the Amazon with complex networks and agent-based modeling. After completing his PhD in 2018, he joined MCC Berlin as a postdoctoral researcher and now uses natural language processing and network analysis to better understand climate politics. [Müller-Hansen et al., 2019, Müller-Hansen et al., 2017b, Müller-Hansen et al., 2017a]

Jan Nitzbon wrote his master's thesis in physics with *copan* between 2015 and 2016 in which he explored pathways of global human-nature coevolution using the copan:GLOBAL model. He has recently obtained his PhD in geography for a thesis investigating permafrost degradation under climate warming using numerical modelling. [Nitzbon et al., 2017]

Ilona M. Otto worked as a post-doctoral researcher in *copan* from 2015 to 2020. She is currently working as a Professor for Societal Impacts of Climate Change and leading a research group Social Complexity and System Transformation at the Wegener Center for Climate and Global Change, University of Graz, Austria.

Erik Scharwächter did his computer science master's thesis in *copan* in 2015, where he studied and developed algorithms to learn evolution rules for social networks. [Scharwächter et al., 2016]

Antonia Schuster is interested in socio-metabolic classes and their mapping to different levels of human agency to improve our understanding of social complexities. As a doctoral researcher, she is part of the *copan* group since 2020. [Schuster and Otto, 2021]

Felix Strnad did his physics master's thesis in *copan* in 2019 in which he dealt with the coupling of deep reinforcement learning and World-Earth modeling. Since September 2020, he works as a PhD student at the International Max Planck Research School for Intelligent Systems (IMPRS-IS) in Tübingen. [Strnad et al., 2019]

Benedikt Stumpf did his master's thesis in physics in *copan* in 2018/2019, working on critical thresholds of tipping cascades on complex networks. He is currently working as a teacher for physics, history and political education. [Krönke et al., 2020, Wunderling et al., 2020d]

Lea Tamberg wrote her Bachelor's thesis in systems science in *copan* in 2020, investigating the effects of no-growth policies on the copan:GLOBAL model. She is currently pursuing a Master's degree in Data Science at ETH Zurich. [Tamberg et al., 2020]

Marc Wiedermann did his physics master's thesis with *copan* in 2014, developing a socio-ecological model of coevolutionary network dynamics. After his PhD he re-joined *copan* as a PostDoc in 2017, working on low-dimensional models for social tipping processes in response to anticipated and experienced climate impacts.

Ricarda Winkelmann, Professor of Climate System Analysis at PIK and University of Potsdam, is an associated member of *copan* , working closely together with Jonathan Donges, Jobst Heitzig and others on interacting tipping elements in the Earth system.

Valentin Wohlfarth did his physics master's thesis in *copan* 2020-2021, working on tipping cascades on the international trade network and the effects of network structures on tipping dynamics. [Wunderling et al., 2021]

Nico Wunderling did his PhD thesis in *copan* from 2017–2021, working on the the emergence of tipping cascades among climate tipping elements under global warming. [Wunderling et al., 2021, Wunderling et al., 2020a, Wunderling et al., 2020b, Wunderling et al., 2020c, Wunderling et al., 2020d, Wunderling et al., 2020e]

Furthermore, Sara Ansari, Sabine Auer, Reik Donner, Vera Heck, Sarah Hiller, Volker Karle, Jan Kohler, Paul Müller and Kilian Zimmerer have contributed to a number of further *copan* publications [Heitzig and Hiller, 2020, Müller et al., 2020, Ansari et al., 2021].

Copan software

Abacra-Model An agent-based Amazonian cattle ranching model. `https://github.com/fmhansen/abacra`. [Müller-Hansen et al., 2019]

CyExploit A cython / python implementation of the copan:EXPLOIT model. `https://github.com/wbarfuss/cyexploit`. [Wiedermann et al., 2015, Barfuss et al., 2017]

EvoMine Algorithm for mining frequently occurring graph evolution rules. `https://hpi.de/mueller/evomine.html`. [Scharwächter et al., 2016]

PyCascades Python framework for simulating tipping cascades on complex networks. `https://github.com/pik-copan/pycascades/tree/v1.0`. [Wunderling et al., 2021]

PyCopanBehave A python implementation of copan:BEHAVE model. `https://github.com/pik-copan/pycopanbehave`. [Schleussner, C. F. and Donges, J. F. et al., 2016]

PyCopanCore A reference implementation of the copan:CORE open World-Earth modelling framework. `https://github.com/pik-copan/pycopancore`. [Donges, J. F. and Heitzig, J. et al., 2020]

PyCopanPbcc Python scripts for modelling collateral transgression of planetary boundaries. `https://github.com/pik-copan/pycopanpbcc`. [Heck et al., 2016]

PyDRLinWESM A package for using Deep Reinforcement Learning within World-Earth Models to discover sustainable management strategies. `https://github.com/fstrnad/pyDRLinWESM`. [Strnad et al., 2019]

PyMofa A collection of simple functions to run and evaluate computer models systematically. `https://github.com/jakobkolb/pymofa`. [Barfuss, 2019]

PyRegimeShifts Python scripts for detecting regime shifts in paleoclimate time series. `https://github.com/pik-copan/pyregimeshifts`. [Donges et al., 2015a]

PyTPT Implementation of Transition Path Theory for: stationary Markov chains, periodically varying Markov chains and time-inhomogenous Markov chains over finite time intervals. `https://github.com/LuzieH/pytpt`. [Helfmann et al., 2020]

PyUnicorn Python modules for complex network and nonlinear time series analysis. `https://github.com/pik-copan/pyunicorn`. [Donges et al., 2015b]

PyViability A library for computations related to viability theory, in particular the viability kernel and the capture basin, and for the classifications of models with respect to the Topology of Sustainable Management. `https://github.com/timkittel/PyViability`. [Kittel et al., 2017b]

Copan references

[Arndt et al., 2016] Arndt, P. S., Donges, J. F., and Heitzig, J. (2016). The Great Acceleration: timing a high-potential candidate for responsibly defining the onset of the Anthropocene. (in review).

[Asano et al., 2019] Asano, Y. M., Kolb, J. J., Heitzig, J., and Doyne Farmer, J. (2019). Emergent inequality and endogenous dynamics in a simple behavioral macroeconomic model. *Proceedings of the National Academy of Sciences of the United States of America*, in press (arXiv:1907.02155).

[Auer et al., 2015] Auer, S., Heitzig, J., Kornek, U., Schöll, E., and Kurths, J. (2015). The Dynamics of Coalition Formation on Complex Networks. *Scientific Reports*, 5:1–8.

[Barfuss, 2019] Barfuss, W. (2019). *Learning dynamics and decision paradigms in social-ecological dilemmas*. PhD thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät.

[Barfuss et al., 2017] Barfuss, W., Donges, F. J., Wiedermann, M., and Lucht, W. (2017). Sustainable use of renewable resources in a stylized social-ecological network model under heterogeneous resource distribution. *Earth System Dynamics*, 8(2):255–264.

[Barfuss et al., 2019] Barfuss, W., Donges, J. F., and Kurths, J. (2019). Deterministic limit of temporal difference reinforcement learning for stochastic games. *Physical Review E*, 99(4):1–16.

[Barfuss et al., 2018] Barfuss, W., Donges, J. F., Lade, S. J., and Kurths, J. (2018). When optimization for governing human-environment tipping elements is neither sustainable nor safe. *Nature Communications*, 9(1):1–10.

[Barfuss et al., 2020] Barfuss, W., Donges, J. F., Vasconcelos, V. V., Kurths, J., and Levin, S. A. (2020). Caring for the future can turn tragedy into comedy for long-term collective action under risk of collapse. *Proceedings of the National Academy of Sciences of the United States of America*, 117(23):12915–12922.

[Donges and Barfuss, 2017] Donges, J. F. and Barfuss, W. (2017). From Math to Metaphors and Back Again Social-Ecological Resilience from a Multi-Agent-Environment Perspective Jonathan. *GAIA*, 26:182–190.

[Donges et al., 2015a] Donges, J. F., Donner, R. V., Marwan, N., Breitenbach, S. F., Rehfeld, K., and Kurths, J. (2015a). Non-linear regime shifts in Holocene Asian monsoon variability: Potential impacts on cultural change and migratory patterns. *Climate of the Past*, 11(5):709–741.

[Donges et al., 2015b] Donges, J. F., Heitzig, J., Beronov, B., Wiedermann, M., Runge, J., Feng, Q. Y., Tupikina, L., Stolbova, V., Donner, R. V., Marwan, N., Dijkstra, H. A., and Kurths, J. (2015b). Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package. *Chaos*, 25(11).

[Donges et al., 2018] Donges, J. F., Lucht, W., Heitzig, J., Barfuss, W., Cornell, S. E., Lade, S. J., and Schlüter, M. (2018). Taxonomies for structuring models for World-Earth system analysis of the Anthropocene: subsystems, their interactions and social-ecological feedback loops. *Earth System Dynamics Discussions*, 2018:1–30.

[Donges et al., 2017] Donges, J. F., Lucht, W., Müller-Hansen, F., and Steffen, W. (2017). The technosphere in Earth System analysis: A coevolutionary perspective. *The Anthropocene ReviewAnthropocene Review*, 4(1):23–33.

[Donges et al., 2016] Donges, J. F., Schleussner, C. F., Siegmund, J. F., and Donner, R. V. (2016). Event coincidence analysis for quantifying statistical interrelationships between event time series: On the role of flood events as triggers of epidemic outbreaks. *The European Physical Journal: Special Topics*, 225(3):471–487.

[Donges, J. F. and Heitzig, J. et al., 2020] Donges, J. F. and Heitzig, J., Barfuss, W., Wiedermann, M., Kassel, J. A., Kittel, T., Kolb, J. J., Kolster, T., Müller-Hansen, F., Otto, I. M., Zimmerer, K. B., and Lucht, W. (2020). Earth system modeling with endogenous and dynamic human societies: The copan:CORE open World-Earth modeling framework. *Earth System Dynamics*, 11(2):395–413.

[Donges, J. F. and Lochner, J. et al., 2021] Donges, J. F. and Lochner, J., Heitzig, J., Kitzmann, N., Lehmann, S., Wiedermann, M., and Vollmer, J. (2021). Dose-response function approach for detecting spreading processes in temporal network data. *arXiv*, (id:2103.09496).

[Donges, J. F. and Winkelmann, R. et al., 2017] Donges, J. F. and Winkelmann, R., Lucht, W., Cornell, S. E., Dyke, J. G., Rockström, J., Heitzig, J., and Schellnhuber, H. J. (2017). Closing the loop: Reconnecting human dynamics to Earth System science. *Anthropocene Review*, 4(2):151–157.

[Ewers, B. and Donges, J. F. et al., 2019] Ewers, B. and Donges, J. F., Heitzig, J., and Peterson, S. (2019). Divestment may burst the carbon bubble if investors' beliefs tip to anticipating strong future climate policy. *arXiv*, (id:1902.07481).

[Geier et al., 2019] Geier, F., Barfuss, W., Wiedermann, M., Kurths, J., and Donges, J. F. (2019). The physics of governance networks: critical transitions in contagion dynamics on multilayer adaptive networks with application to the sustainable use of renewable resources. *European Physical Journal: Special Topics*, 228(11):2357–2369.

[Heck et al., 2016] Heck, V., Donges, J. F., and Lucht, W. (2016). Collateral transgression of planetary boundaries due to climate engineering by terrestrial carbon dioxide removal. *Earth System Dynamics*, 7(4):783–796.

[Heitzig et al., 2018] Heitzig, J., Barfuss, W., and Donges, J. F. (2018). A thought experiment on sustainable management of the earth system. *Sustainability*, 10(6):1–25.

[Heitzig and Hiller, 2020] Heitzig, J. and Hiller, S. (2020). Degrees of individual and groupwise backward and forward responsibility in extensive-form games with ambiguity, and their application to social choice problems. *arXiv*, (id:2007.07352).

[Heitzig et al., 2016] Heitzig, J., Kittel, T., Donges, J. F., and Molkenthin, N. (2016). Topology of sustainable management of dynamical systems with desirable states: From defining planetary boundaries to safe operating spaces in the Earth system. *Earth System Dynamics*, 7(1):21–50.

[Heitzig and Kornek, 2018] Heitzig, J. and Kornek, U. (2018). Bottom-up linking of carbon markets under far-sighted cap coordination and reversibility. *Nature Climate Change*, 8(3):204–209.

[Helfmann et al., 2021] Helfmann, L., Conrad, N. D., Djurdjevac, A., Winkelmann, S., and Schütte, C. (2021). From Interacting Agents To Density-Based Modeling With Stochastic Pdes. *Communications in Applied Mathematics and Computational Science*, 16(1):1–32.

[Helfmann et al., 2020] Helfmann, L., Ribera Borrell, E., Schütte, C., and Koltai, P. (2020). Extending Transition Path Theory: Periodically Driven and Finite-Time Dynamics. *Journal of Nonlinear Science*, 30(6):3321–3366.

[Holstein et al., 2021] Holstein, T., Wiedermann, M., and Kurths, J. (2021). Optimization of coupling and global collapse in diffusively coupled socio-ecological resource exploitation networks. *New Journal of Physics*, 23(3):033027.

[Kittel et al., 2017a] Kittel, T., Heitzig, J., Webster, K., and Kurths, J. (2017a). Timing of transients: Quantifying reaching times and transient behavior in complex systems. *New Journal of Physics*, 19(8).

[Kittel et al., 2017b] Kittel, T., Müller-Hansen, F., Koch, R., Heitzig, J., Deffuant, G., Mathias, J.-D., and Kurths, J. (2017b). From lakes and glades to viability algorithms: Automatic classification of system states according to the Topology of Sustainable Management. *arXiv*, (id:1706.04542).

[Klamser et al., 2017] Klamser, P. P., Wiedermann, M., Donges, J. F., and Donner, R. V. (2017). Zealotry effects on opinion dynamics in the adaptive voter model. *Physical Review E*, 96(5).

[Klose et al., 2020] Klose, A. K., Karle, V., Winkelmann, R., and Donges, J. F. (2020). Emergence of cascading dynamics in interacting tipping elements of ecology and climate. *Royal Society Open Science*, 7(6):200599.

[Kolb et al., 2020] Kolb, J. J., Müller-Hansen, F., Kurths, J., and Heitzig, J. (2020). Macroscopic approximation methods for the analysis of adaptive networked agent-based models: Example of a two-sector investment model. *Physical Review E*, 102(4).

[Krönke et al., 2020] Krönke, J., Wunderling, N., Winkelmann, R., Staal, A., Stumpf, B., Tuinenburg, O. A., and Donges, J. F. (2020). Dynamics of tipping cascades on complex networks. *Physical Review E*, 101(4):1–19.

[Müller et al., 2020] Müller, P. M., Heitzig, J., Kurths, J., Lüdge, K., and Wiedermann, M. (2020). Anticipation-induced social tipping - Can the environment be stabilised by social dynamics? *European Physical Journal: Special Topics*, in press(arXiv:2012.01977).

[Müller-Hansen et al., 2017a] Müller-Hansen, F., Cardoso, M. F., Dalla-Nora, E. L., Donges, J. F., Heitzig, J., Kurths, J., and Thonicke, K. (2017a). A matrix clustering method to explore patterns of land-cover transitions in satellite-derived maps of the Brazilian Amazon. *Nonlinear Processes in Geophysics*, 24(1):113–123.

[Müller-Hansen et al., 2019] Müller-Hansen, F., Heitzig, J., Donges, J. F., Cardoso, M. F., Dalla-Nora, E. L., Andrade, P., Kurths, J., and Thonicke, K. (2019). Can Intensification of Cattle Ranching Reduce Deforestation in the Amazon? Insights From an Agent-based Social-Ecological Model. *Ecological Economics*, 159:198–211.

[Müller-Hansen et al., 2017b] Müller-Hansen, F., Schlüter, M., Mäs, M., Donges, J. F., Kolb, J. J., Thonicke, K., and Heitzig, J. (2017b). Towards representing human behavior and decision making in Earth system models - An overview of techniques and approaches. *Earth System Dynamics*, 8(4):977–1007.

[Nitzbon et al., 2017] Nitzbon, J., Heitzig, J., and Parlitz, U. (2017). Sustainability, collapse and oscillations in a simple World-Earth model. *Environmental Research Letters*, 12(7).

[Otto et al., 2015] Otto, I. M., Biewald, A., Coumou, D., Feulner, G., Köhler, C., Nocke, T., Blok, A., Gröber, A., Selchow, S., Tyfield, D., Volkmer, I., Schellnhuber, H. J., and Beck, U. (2015). Socio-economic data for global environmental change research. *Nature Climate Change*, 5(6):503–506.

[Otto et al., 2020a] Otto, I. M., Donges, J. F., Cremades, R., Bhowmik, A., Hewitt, R. J., Lucht, W., Rockström, J., Allerberger, F., McCaffrey, M., Doe, S. S., Lenferna, A., Morán, N., van Vuuren, D. P., and Schellnhuber, H. J. (2020a). Social tipping dynamics for stabilizing Earth's climate by 2050. *Proceedings of the National Academy of Sciences of the United States of America*, 117(5):2354–2365.

[Otto et al., 2019] Otto, I. M., Kim, K. M., Dubrovsky, N., and Lucht, W. (2019). Shift the focus from the super-poor to the super-rich. *Nature Climate Change*, 9(2):82–84.

[Otto et al., 2020b] Otto, I. M., Wiedermann, M., Cremades, R., Donges, J. F., Auer, C., and Lucht, W. (2020b). Human agency in the Anthropocene. *Ecological Economics*, 167.

[Scharwächter et al., 2016] Scharwächter, E., Müller, E., Donges, J., Hassani, M., and Seidl, T. (2016). Detecting change processes in dynamic networks by frequent graph evolution rule mining. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1191–1196.

[Schleussner et al., 2016] Schleussner, C. F., Donges, J. F., Donner, R. V., and Schellnhuber, H. J. (2016). Armed-conflict risks enhanced by climate-related disasters in ethnically fractionalized countries. *Proceedings of the National Academy of Sciences of the United States of America*, 113(33):9216–9221.

[Schleussner, C. F. and Donges, J. F. et al., 2016] Schleussner, C. F. and Donges, J. F., Engemann, D. A., and Levermann, A. (2016). Clustered marginalization of minorities during social transitions induced by co-evolution of behaviour and network structure. *Scientific Reports*, 6(1):30790.

[Schuster and Otto, 2021] Schuster, A. and Otto, I. M. (2021). Socio-metabolic class conflicts in the Anthropocene: Developing a novel class theory based on German population data. *Capitalism Nature Socialism*, in press.

[Strnad et al., 2019] Strnad, F. M., Barfuss, W., Donges, J. F., and Heitzig, J. (2019). Deep reinforcement learning in World-Earth system models to discover sustainable management strategies. *Chaos*, 29(12):123122.

[Tamberg et al., 2020] Tamberg, L. A., Heitzig, J., and Donges, J. F. (2020). A modeler's guide to studying the resilience of social-technical-environmental systems. *arXiv*, (id:2007.05769).

[Van Kan et al., 2016] Van Kan, A., Jegminat, J., Donges, J. F., and Kurths, J. (2016). Constrained basin stability for studying transient phenomena in dynamical systems. *Physical Review E*, 93(4):1–7.

[Wiedermann et al., 2015] Wiedermann, M., Donges, J. F., Heitzig, J., Lucht, W., and Kurths, J. (2015). Macroscopic description of complex adaptive networks co-evolving with dynamic node states. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 91(5):1–11.

[Wiedermann et al., 2020] Wiedermann, M., Smith, E. K., Heitzig, J., and Donges, J. F. (2020). A network-based microfoundation of Granovetter's threshold model for social tipping. *Scientific Reports*, 10(1):1–10.

[Winkelmann, R. and Donges, J. F. and Smith, E. K. and Milkoreit, M. et al., 2020] Winkelmann, R. and Donges, J. F. and Smith, E. K. and Milkoreit, M., Eder, C., Heitzig, J., Katsanidou, A., Wiedermann, M., Wunderling, N., and Lenton, T. M. (2020). Social tipping processes for sustainability: An analytical framework. *arXiv*, (doi:10.2139/ssrn.3708161).

[Wunderling et al., 2020a] Wunderling, N., Donges, J., Kurths, J., and Winkelmann, R. (2020a). Interacting tipping elements increase risk of climate domino effects under global warming. *Earth System Dynamics Discussions*, 2020(April):1–21.

[Wunderling et al., 2020b] Wunderling, N., Gelbrecht, M., Winkelmann, R., Kurths, J., and Donges, J. F. (2020b). Basin stability and limit cycles in a conceptual model for climate tipping cascades. *New Journal of Physics*, 22:123031.

[Wunderling et al., 2021] Wunderling, N., Krönke, J., Wohlfarth, V., Kohler, J., Heitzig, J., Staal, A., Willner, S., Winkelmann, R., and Donges, J. F. (2021). Modelling nonlinear dynamics of interacting tipping elements on complex networks: the PyCascades package. *European Physical Journal: Special Topics*, in press(arXiv:2011.02031).

[Wunderling et al., 2020c] Wunderling, N., Staal, A., Sakschewski, B., Hirota, M., Tuinenburg, O., Donges, J., Barbosa, H., and Winkelmann, R. (2020c). Network dynamics of drought-induced tipping cascades in the Amazon rainforest. *Research Square*, (doi:10.21203/rs.3.rs-71039/v1).

[Wunderling et al., 2020d] Wunderling, N., Stumpf, B., Krönke, J., Staal, A., Tuinenburg, O. A., Winkelmann, R., and Donges, J. F. (2020d). How motifs condition critical thresholds for tipping cascades in complex networks: Linking micro- to macro-scales. *Chaos*, 30(4):043129.

[Wunderling et al., 2020e] Wunderling, N., Willeit, M., Donges, J. F., and Winkelmann, R. (2020e). Global warming due to loss of large ice masses and Arctic summer sea ice. *Nature Communications*, 11(1).

Further References

[Ansari et al., 2021] Ansari, S., Anvari, M., Pfeffer, O., Molkenthin, N., Hellmann, F., Heitzig, J., and Kurths, J. (2021). Moving the epidemic tipping point through topologically targeted social distancing. *arXiv*, (id:2102.09997).

[Farmer et al., 2019] Farmer, J. D., Hepburn, C., Ives, M. C., Hale, T., Wetzer, T., Mealy, P., Rafaty, R., Srivastav, S., and Way, R. (2019). Sensitive intervention points in the post-carbon transition. *Science*, 364(6436):132–134.

[Hellmann et al., 2016] Hellmann, F., Schultz, P., Grabow, C., Heitzig, J., and Kurths, J. (2016). Survivability of Deterministic Dynamical Systems. *Scientific Reports*, 6(29654):1–12.

[Holling, 1973] Holling, C. S. (1973). Resilience and stability of ecological systems. *The Future of Nature: Documents of Global Change*, 4:1–23.

[Holme and Newman, 2006] Holme, P. and Newman, M. E. (2006). Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74(5):1–5.

[Lade et al., 2017] Lade, S. J., Bodin, Ö., Donges, J. F., Kautsky, E. E., Galafassi, D., Olsson, P., and Schlüter, M. (2017). Modelling social-ecological transformations: An adaptive network proposal. *arXiv*, (id:1704.06135).

[Lenton, 2020] Lenton, T. M. (2020). Tipping positive change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1794):20190123.

[Menck et al., 2013] Menck, P. J., Heitzig, J., Marwan, N., and Kurths, J. (2013). How basin stability complements the linear-stability paradigm. *Nature Physics*, 9(2):89–92.

[Milkoreit et al., 2018] Milkoreit, M., Hodbod, J., Baggio, J., Benessaiah, K., Calderón-Contreras, R., Donges, J. F., Mathias, J. D., Rocha, J. C., Schoon, M., and Werners, S. E. (2018). Defining tipping points for social-ecological systems scholarship - An interdisciplinary literature review. *Environmental Research Letters*, 13(3).

[Petschel-Held et al., 1999] Petschel-Held, G., Block, A., Cassel-Gintz, M., Kropp, J., Lüdeke, M. K., Moldenhauer, O., Reusswig, F., and Schellnhuber, H. J. (1999). Syndromes of Global Change: A qualitative modelling approach to assist global environmental management. *Environmental Modeling and Assessment*, 4(4):295–314.

[Raworth, 2017] Raworth, K. (2017). A Doughnut for the Anthropocene: humanity's compass in the 21st century. *The Lancet Planetary Health*, 1(2):e48–e49.

[Rockström et al., 2017] Rockström, J., Gaffney, O., Rogelj, J., Meinshausen, M., Nakicenovic, N., and Schellnhuber, H. J. (2017). A roadmap for rapid decarbonization.

[Rockström et al., 2009] Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R. W., Fabry, V. J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., and Foley, J. (2009). Planetary boundaries: Exploring the safe operating space for humanity. *Ecology and Society*, 14(2).

[Schellnhuber, 1998] Schellnhuber, H. J. (1998). Discourse: Earth System Analysis - The Scope of the Challenge. In *Earth System Analysis*, pages 3–195.

[Steffen et al., 2018] Steffen, W., Rockström, J., Richardson, K., Lenton, T. M., Folke, C., Liverman, D., Summerhayes, C. P., Barnosky, A. D., Cornell, S. E., Crucifix, M., Donges, J. F., Fetzer, I., Lade, S. J., Scheffer, M., Winkelmann, R., and Schellnhuber, H. J. (2018). Trajectories of the Earth System in the Anthropocene. *Proceedings of the National Academy of Sciences of the United States of America*, 115(33):8252–8259.