POTSDAM-INSTITUT FÜR
KLIMAFOLGENFORSCHUNG

Bottom-Up Linking of Carbon Markets Under Farsighted Cap Coordination and Reversibility

Jobst Heitzig*
Potsdam Institute for Climate Impact Research,
Transdisciplinary Concepts and Methods,
P. O. Box 60 12 03, 14412 Potsdam, Germany
heitzig@pik-potsdam.de

Ulrike Kornek
Mercator Research Institute on Global Commons and Climate Change,
EUREF Campus 19, Torgauer Str. 12–15, 10829 Berlin, Germany
kornek@mcc-berlin.net

Manuscript for Nature Climate Change
Received: date / Accepted: date

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 since [4, 5] finds meager prospects for effective collaboration [6]. Countries also face incentives to increase emissions when linking their markets without a cap agreement [7, 8]. Here, we analyze the dynamics of market linkage using a game-theoretic model of farsighted coalition formation. In contrast to non-dynamic models and dynamic models without farsightedness [9, 10], in our model an efficient global coalition always forms eventually if players are sufficiently farsighted or caps are coordinated immediately when markets are linked.

Our study extends the climate coalition literature by analyzing a *dynamic* process with *multiple* coalitions, *farsighted* players anticipating further steps, and *uncertainty* about which transitions will happen [11, 12, 13] (in contrast to cost and benefit uncertainty). We adapt, to our knowledge for the first time, the dynamic farsighted coalition formation model of [13] to the linking of carbon markets with endogenous decisions whether to coordinate caps. Unlike [14, 15, 16, 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 (e.g., [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 are 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 (e.g., 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 (e.g., 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, later a version of the coalitional payoffs from [7] based on cost-benefit estimates from [23, 24], assuming that surplusses from forming a coalition are shared according to the asymmetric Nash bargaining solution [25], i.e., in proportion to some distribution of *bargaining power,* see *Methods: Derivation of static payoffs.*

The possible *transitions* between states $x, y, \ldots$ (arrows in Fig. 1) represent the formation of new markets or coalitions (e.g., 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 (e.g., 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 *farsightedness* δ used in evaluations, representing the combined effects of time preference, trust that the process does not break down, and duration between steps. Mathematically,

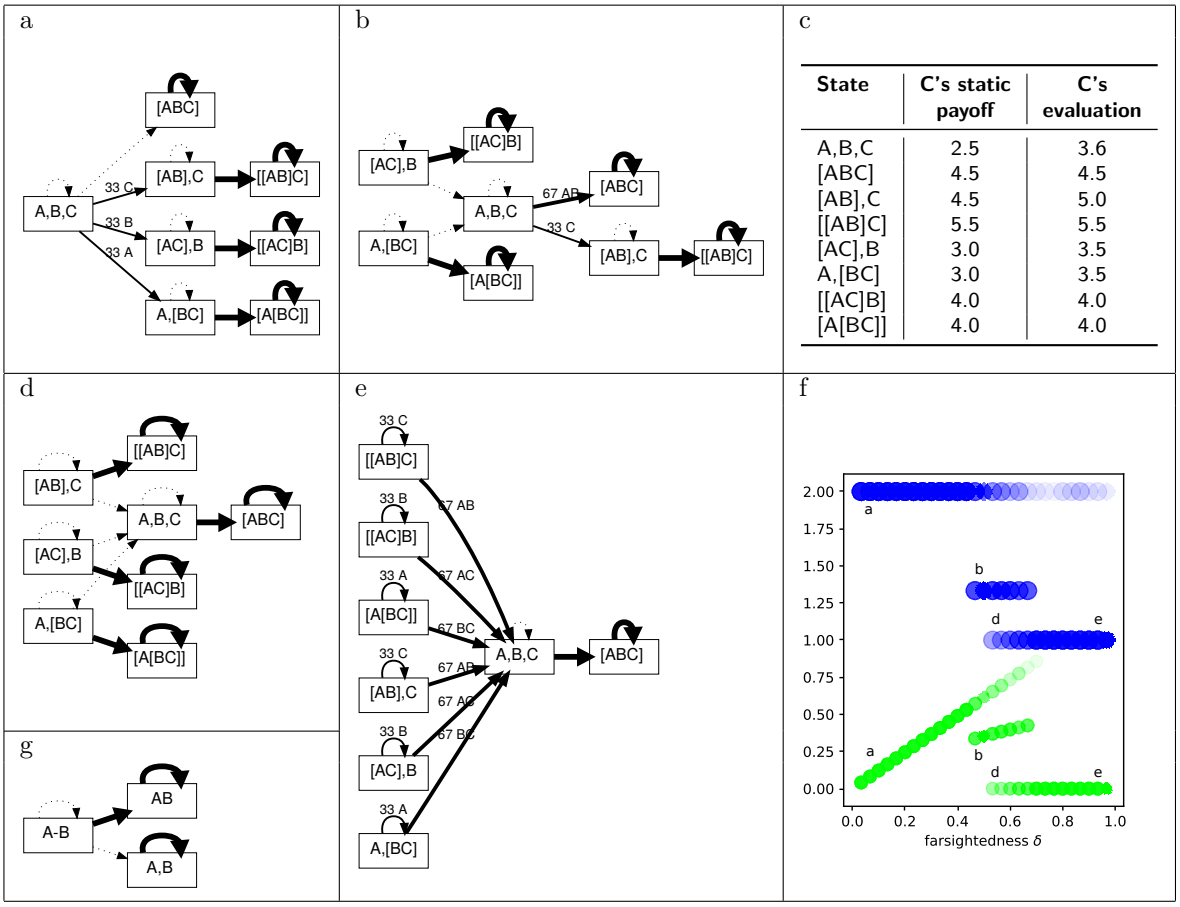
$$\ell_i(x) = (1-\delta)\pi_i(x) + \delta \sum_y p_{x\to y}\ell_i(y). \quad (1)$$

At the same time, given any such evaluations, certain transitions appear *unprofitable* since they decrease the evaluation of some relevant player (e.g., 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 accept this since they profit from it and cannot initiate a better transition. Note that she may propose a coalition that excludes herself (e.g., 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,

$$\{p_{x\to y}\} = f(\{\ell_i(x)\}), \quad (2)$$

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 w.r.t. 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.



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

Figure 1: Illustration of the model in a fictitious situation with three symmetric players A,B,C, eight possible coalition states, linear mitigation benefits, and quadratic mitigation costs. (a) unique equilibrium process for low farsightedness $\delta = 0.3$; arrow labels state transition probabilities in percent and which players favour this transition. (b) one of three alternative equilibrium processes for medium farsightedness $\delta = 0.5$. (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 farsightedness $\delta = 0.7$. (e) very high farsightedness $\delta = 0.9$. (f) effect of farsightedness on mean no. of steps to reach a grand coalition (blue) and on total payoff uncertainty (green, 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 SI 3.6 for details). (g) example where two asymmetric players can get stuck when immediate cap coordination is unavailable ($\delta = 0.5$, see SI 3.2 for details).

Consider the illustrative example of Fig. 1, where three symmetric players can form coalitions with static payoffs based on [26]. Player i's benefits and costs from mitigating q_i units of 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 farsightedness ($\delta < 0.45$) 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 (e.g., C in Fig. 1b,c) would not join a bilateral coalition, resulting in a 2/3 probability

of forming the 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 (e.g., "[[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 farsighted enough to prefer terminating their coalition, anticipating the eventually higher payoffs in "[ABC]" (Fig. 1e). While for each given type of equilibrium, increasing farsightedness 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 SI 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 emitters C(hina), E(urope), F(ormer Soviet Union), I(ndia), J(apan), and U(SA) are derived from the literature ([7, 23, 24], see *Methods*), resulting in scenarios such as those depicted in Figs. 2–4. Table 1 compares the (myopic) static payoffs and (farsighted) 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 would 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.

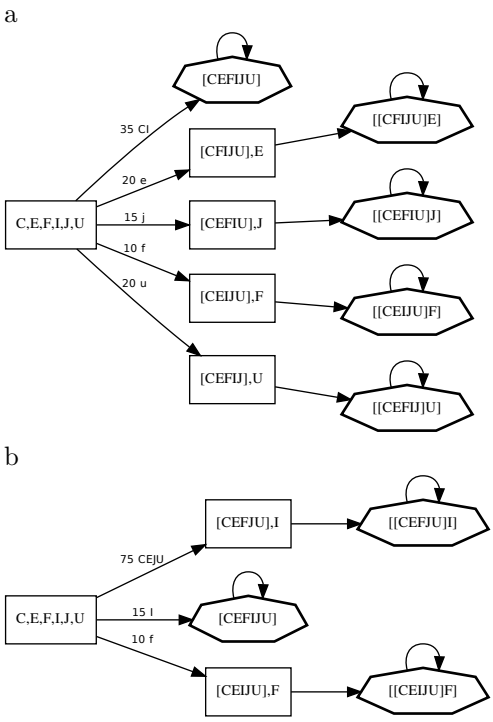


Figure 2: (a) Typical model result for the six major emitters with low to moderate farsightedness (here $\delta = 0.5$). A fully coordinated global carbon market results after one step (with 35% probability, if the permit sellers C(hina) and I(ndia) get their way) or two steps (with 65% probability, 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) Typical result for highly farsighted players.

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 SI 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.

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

Table 1: Static payoffs derived from [7, 23, 24] and evaluations [bln. US$ per 100 years] in the process shown in Fig. 2a for a typical choice of parameters (medium farsightedness $\delta = 0.5$, agreements unilaterally terminable), for states reached with positive probability (boldface) and some alternative states. * Favourite undominated move of this player in state C,E,F,I,J,U. > Move is not statically profitable for this initiating player. ≫ Move is not long-term profitable for this initiating player. (a),(b): states referred to in main text.

- Counter-intuitively, 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 farsightedness tends to reduce uncertainty and mean no. 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 (e.g., the "C-E-F-I-J-U" branch in Fig. 3). Higher farsightedness here tends to *increase* uncertainty and mean no. 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 farsighted, 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. (i) 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 (e.g., 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 (e.g., compare the evaluations for C in state [CEFJU],I in Tables 1 and SI 2, and C's favourite moves in states C,E,F,I,J,U and [CFJU],E,I of Fig. 3). (ii) Free-riding by not coordinating caps: In a not yet fully coordinated market, both permit buyers and sellers usually have an incentive to

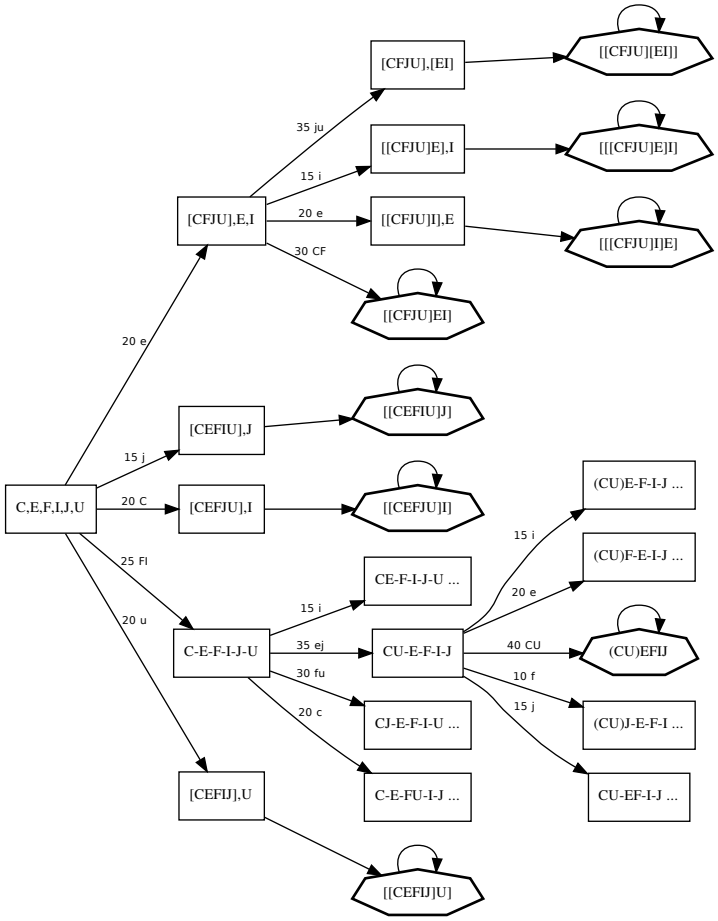


Figure 3: Alternative to Fig. 2ab with typical complications occurring if players are highly farsighted ($\delta = 0.9$) and agreements are irreversible (see Table SI 1 for payoffs and evaluations and Sec. SI 3.3 for a discussion). In view of the expected later moves, F and I now prefer to establish 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 "...").

free-ride by entering a coalition late (e.g., in the "C-E-F-I-J-U" branch in Fig. 3). Overall, the analysis in SI 3.6 shows 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

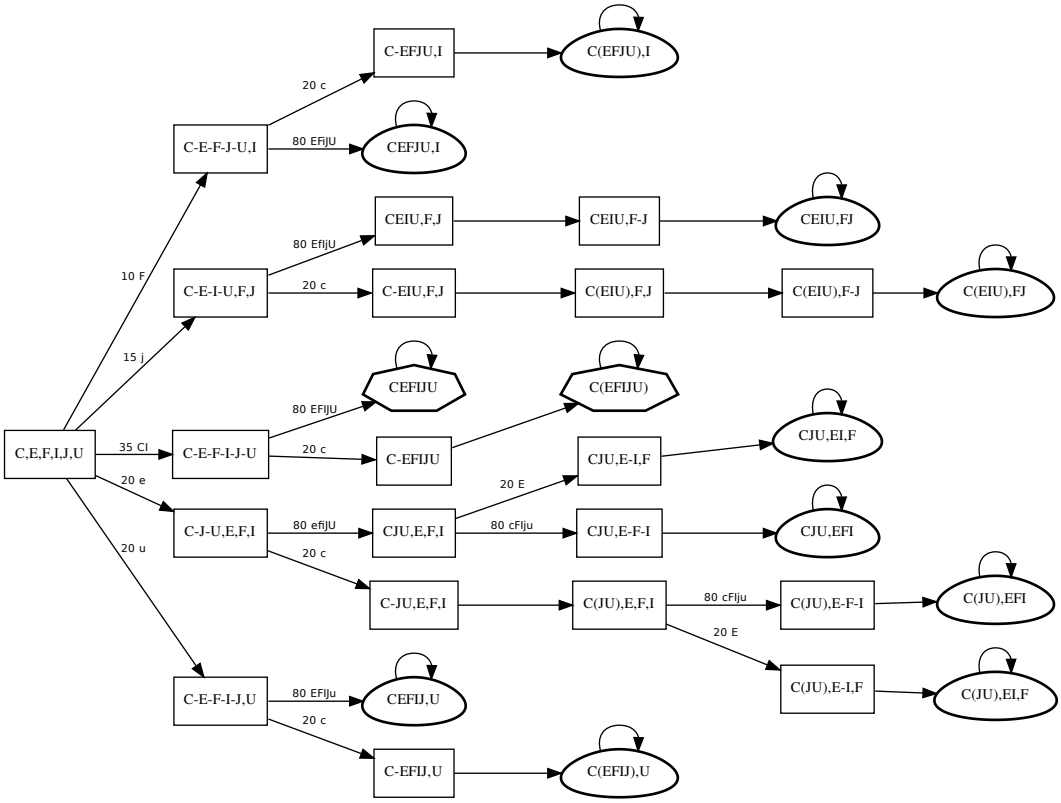


Figure 4: Alternative scenario to Fig. 2 in a world where caps cannot immediately be coordinated when markets are linked but only later in separate moves (medium farsightedness of $\delta = 0.5$, unilaterally terminable agreements; see Figs. SI 3 and SI 4 for irreversible agreements and myopic players). 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 farsighted enough to accept the temporary costs of delayed cap coordination.

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 currently 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 farsightedness, 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 presently only few coordinated carbon markets.

Acknowledgements

The authors thank Kai Lessmann, Robert Marschinski, Ottmar Edenhofer, the Policy Instruments Group and the COPAN Flagship Project at the Potsdam Institute for Climate Impact Research for many intense discussions; Bjart Holtsmark, Mads Greaker, Cathrine Hagem, Robert Schmidt, and the CREW project for inspiring work; and Rajiv Vohra, Peter Menck, Norbert Marwan, Jonathan Donges, and Carl-Friedrich Schleussner for helpful comments.

Author contributions

Jobst Heitzig developed the model and conducted the numerical experiments. Both authors interpreted the study, and wrote and edited the text.

References

[1] UNFCCC. Adoption of the paris agreement. report no. fccc/cp/2015/l.9/rev.1. article 6 (2015). URL `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. *Nature Climate 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. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 15739–15744 (2011).

[4] Carraro, C. & Siniscalco, D. Strategies for the international protection of the environment. *Journal of Public Economics* **52**, 309–328 (1993).

[5] Barrett, S. Self-enforcing international environmental agreements. *Oxford Economic Papers* (1994).

[6] Finus, M. New Developments in Coalition Theory. In Marsiliani, L., Rauscher, M. & Withagen, C. (eds.) *Environmental policy in an international perspective*, 19–49 (Kluwer, 2003).

[7] Helm, C. International emissions trading with endogenous allowance choices. *Journal of Public Economics* **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. *Journal of Environmental Economics and Management* **58**, 266–280 (2009).

[9] Smead, R., Sandler, R. L., Forber, P. & Basl, J. A bargaining game analysis of international climate negotiations. *Nature Climate Change* **4**, 442–445 (2014).

[10] Verendel, V., Johansson, D. J. A. & Lindgren, K. Strategic reasoning and bargaining in catastrophic climate change games. *Nature Climate Change* **6**, 6–10 (2015).

[11] Ray, D. & Vohra, R. Equilibrium Binding Agreements. *Journal of Economic Theory* **73**, 30–78 (1997).

[12] Ray, D. & Vohra, R. A theory of endogenous coalition structures. *Games and Economic Behavior* **26**, 286–336 (1999).

[13] Konishi, H. & Ray, D. Coalition formation as a dynamic process. *Journal of Economic Theory* **110**, 1–41 (2003).

[14] de Zeeuw, A. Dynamic effects on the stability of international environmental agreements. *Journal of Environmental Economics and Management* **55**, 163–174 (2008).

[15] Biancardi, M. & Villani, G. Largest consistent set in international environmental agreements. *Computational Economics* **38**, 407–423 (2011).

[16] Osmani, D. A Note on Computational Aspects of Farsighted 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 Institute for research in economics and business administration Working Papers* **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. *Climate Policy* **9**, 358–372 (2009).

[19] Tuerk, A., Mehling, M., Flachsland, C. & Sterk, W. Linking carbon markets: concepts, case studies and pathways. *Climate 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. *Climate Policy* **16**, 284–300 (2016).

[23] Ellerman, A. D. & Decaux, A. Analysis of Post-Kyoto CO2 Emissions Trading Using Marginal Abatement Curves. *MIT Joint Program on the Science and Policy of Global Change* **Report 40**, 1–33 (1998).

[24] Finus, M., van Ierland, E. & Dellink, R. Stability of Climate Coalitions in a Cartel Formation Game. *Economics of Governance* **7**, 271–291 (2006).

[25] Kalai, E. Nonsymmetric Nash solutions and replications of 2-person bargaining. *International Journal of Game Theory* **6**, 129–133 (1977).

[26] Barrett, S. International Environmental Agreements as Games. In Pethig R. (ed.) *Conflicts and Cooperation in Managing Environmental Resources* (Springer, Berlin, Heidelberg, 1992).

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). E.g., 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 [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 *farsightedness* (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" (i.e., 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 (i.e., 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 (e.g., 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 SI 1.1.2).

See *General game structure* and SI 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 (i.e., a central authority in the respective country or world region, e.g., 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 EU ETS. In the terminology of [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.[1] We represent the *market structure* by a code in which the markets are separated by commas and the members by dashes. E.g., 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* $e_i(t)$

[1] This is justified, e.g., for "two-way direct links" in the terminology of [20] aka "formally linked" markets in the terminology of [21].

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 *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 [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. E.g., 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 *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.[2]

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, i.e., with immediate cap coordination, it is assumed that it can no longer be linked with further markets by the first kind of agreement, i.e., 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 surplusses.

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

$$\begin{array}{c} \text{C-U,E-F-J,I} \\ \xrightarrow{\text{EF}} \\ \text{C-U,EF-J,I} \\ \xrightarrow{\text{EFJ}} \\ \text{C-U,(EF)J,I} \\ \xrightarrow{\text{CU}} \\ \text{CU,(EF)J,I} \\ \xrightarrow{\text{CEFJU}} \\ \text{[(CU)((EF)J)],I} \\ \xrightarrow{\text{CEFIJU}} \\ \text{[[(CU)((EF)J)]I].} \end{array}$$

The number of theoretically possible states grows faster than exponentially in the number of players.

[2] Alternatively, one might enlarge the set of possible states from a finite set to a continuum by representing the state of a market as a pair consisting of a partition of the market's members into coalitions and a set of payoff allocations for these coalitions.

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, e.g., look as depicted in Figs. 2–4 where the arrows are labelled with transition probabilities and those players that favour the move.

Individual and collective rationality, farsightedness. 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 farsighted way, anticipating the further development of the structure. We model the level of this farsightedness via a number $\delta \in (0, 1)$ used in the discounting of prospective future states' static payoffs π_i. This *farsightedness* 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 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 [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.[3]

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,

[3] One might also consider 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 [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.

anticipating its effect on trading (i.e., 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 ([24] version) which calibrates its benefit estimates to the vastly used DICE model of [29] and takes its cost estimates from [23], because that model presents a good trade-off between tractability and qualitative real-world relevance.[4] To keep our numbers comparable to those in [24], we report e_i in Gton CO_2 emissions per 100 years and π_i in bln. US$ 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 $\sigma_i Q$ of global contributions $Q = \sum_{i\in P} q_i = E^0 - E$, and costs of abatement are a cubic function

$$g_i(q_i) = a_i q_i^3/3 + b_i q_i^2/2 \tag{4}$$

of individual contributions, where the coefficients σ_i, a_i, b_i are given in Table SI 1 using calibration I from [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 [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 Fig. SI 1 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

$$\frac{d}{dc_M} p_M = -1 \Big/ \sum_{i\in M} \frac{1}{g_i''(e_i^0 - e_i)} < 0. \tag{8}$$

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

$$\begin{aligned} 0 &= \frac{d}{dc_K}\pi_K \\ &= \sigma_K \frac{d}{dc_K} Q + \sum_{i\in K} [g_i'(e_i^0 - e_i) - p_M] \frac{d}{dc_K} e_i \\ &\quad + p_M + (c_K - e_K)\frac{d}{dc_K} p_M \\ &= p_M - \sigma_K + (c_K - e_K)\frac{d}{dc_M} p_M \end{aligned} \tag{10}$$

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

[4] For future applications, one may use newer estimates, e.g., derived from [30] or from more sophisticated models such as the one in [8].

which can easily be found analytically by summing the above condition over all n_M coalitions, giving

$$\begin{aligned} 0 &= n_M p_M - \sigma_M + (c_M - e_M)\frac{d}{dc_M} p_M \\ &= n_M p_M - \sigma_M \end{aligned} \tag{11}$$

by Eq. 7. Hence the market price is simply

$$p_M = \sigma_M / n_M, \tag{12}$$

actual individual emissions are

$$\begin{aligned} e_i &= e_i^0 - (g_i')^{-1}(p_M) \\ &= e_i^0 - \frac{\sqrt{b_i^2 + 4a_i p_M} - b_i}{2a_i} \end{aligned} \tag{13}$$

by Eq. 6, the coalition's cap choice is

$$c_K = e_K + (p_M - \sigma_K) \sum_{i \in M} \frac{1}{2a_i(e_i^0 - e_i)}, \tag{14}$$

by Eqs. 8, 10, and 13, and all coalitions' payoffs are given by Eq. 9.

From this general payoff structure, [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.

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. E.g., 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,[5] so that

$$\pi_i = \pi_i^0 + \Delta\pi_K w_i / \sum_{j \in K} w_j. \tag{15}$$

For player's bargaining power weights, we use a subjectively chosen distribution that aims at a simple compromise between the following possible choices (see Table SI 1):

- w_i = population of i.
- w_i = GDP of i in US$.
- $w_i = \sigma_i$ (climate "vulnerability").
- $w_i = 1$ (equal bargaining power).
- An "egalitarian" approach that leads to equal per-capita surplus in purchasing power parities (PPP):

$$\begin{aligned} w_i = (&\text{population of } i) \\ &\times (\text{PPP in currency of } i) \\ &\times (\text{exchange rate from } i \text{ to US\$}). \end{aligned}$$

[5] 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 [32] who 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.

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.

Methods References

[27] Heitzig, J. Efficiency in face of externalities when binding hierarchical agreements are possible. *Game Theory & Bargaining 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 and Welfare* **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. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 11721–11726 (2010).

[31] Kalai, E. Nonsymmetric Nash solutions and replications of 2-person bargaining. *International Journal of Game Theory* **6**, 129–133 (1977).

[32] Kalai, E. & Samet, D. On weighted Shapley values. *International Journal of Game Theory* **16**, 205–222 (1987).