ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Integrated assessment model diagnostics: key indicators and model evolution

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
23 December 2020

REVISED
25 March 2021

ACCEPTED FOR PUBLICATION
19 April 2021

PUBLISHED
10 May 2021

LETTER

Integrated assessment model diagnostics: key indicators and model evolution

Mathijs Harmsen[1,2], Elmar Kriegler[3,21], Detlef P van Vuuren[1,2], Kaj-Ivar van der Wijst[1,2], Gunnar Luderer[3,20], Ryna Cui[4], Olivier Dessens[5], Laurent Drouet[6], Johannes Emmerling[6], Jennifer Faye Morris[7], Florian Fosse[8], Dimitris Fragkiadakis[9], Kostas Fragkiadakis[9], Panagiotis Fragkos[9], Oliver Fricko[10], Shinichiro Fujimori[11], David Gernaat[1,2], Céline Guivarch[12], Gokul Iyer[13], Panagiotis Karkatsoulis[9], Ilkka Keppo[14], Kimon Keramidas[8], Alexandre Köberle[15], Peter Kolp[10], Volker Krey[10], Christoph Krüger[1,2], Florian Leblanc[12], Shivika Mittal[15], Sergey Paltsev[7], Pedro Rochedo[16], Bas J van Ruijven[10], Ronald D Sands[17], Fuminori Sano[18], Jessica Strefler[3], Eveline Vasquez Arroyo[16], Kenichi Wada[18] and Behnam Zakeri[10,19]

[1] PBL Netherlands Environmental Assessment Agency, Bezuidenhoutseweg 30, 2594 AV The Hague, The Netherlands
[2] Copernicus Institute for Sustainable Development, Utrecht University, Princetonlaan 8a, 3584 CB Utrecht, The Netherlands
[3] Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam D-14412, Germany
[4] Center for Global Sustainability, University of Maryland, 3101 Van Munching Hall, College Park, MD 20742, United States of America
[5] University College London, London, United Kingdom
[6] RFF-CMCC European Institute on Economics and the Environment (EIEE), Centro Euro-Mediterraneo sui Cambiamenti Climatici, Via Bergogne 34, 20144 Milan, Italy
[7] MIT Joint Program on the Science and Policy of Global Change, Massachusetts Institute of Technology, Cambridge, MA, United States of America
[8] European Commission, Joint Research Centre, Seville, Spain
[9] E3Modelling S.A., Panormou 70-72, Athens, Greece
[10] International Institute for Applied Systems Analysis, Schlossplatz-1, A-2361 Laxenburg, Austria
[11] Department of Environmental Engineering, Kyoto University, Kyoto, Japan & National Institute for Environmental Studies, Center for Social and Environmental Systems Research, Tsukuba, Ibaraki 305-8506, Japan
[12] Ecole des Ponts ParisTech, CIRED, 45bis avenue de la Belle Gabrielle, Nogent-sur-Marne, France
[13] Joint Global Change Research Institute, Pacific Northwest National Laboratory and University of Maryland, 5825 University Research Court, Suite 3500, College Park, MD 20740, United States of America
[14] Department of Mechanical Engineering, School of Engineering, Aalto University, Otakaari 4, Espoo 02150, Finland
[15] Grantham Institute, Imperial College London, Exhibition Road, London SW7 2AZ
[16] Energy Planning Program, COPPE, Universidade Federal do Rio de Janeiro (UFRJ), PO Box 68565, 21941-914 Rio de Janeiro, RJ, Brazil
[17] USDA Economic Research Service, Kansas City, MO, United States of America
[18] Research Institute of Innovative Technology for the Earth (RITE), 9-2, Kizugawadai, Kizugawa-Shi, Kyoto 619-0292, Japan
[19] Sustainable Energy Planning Research Group, Aalborg University, A. C. Meyers Vnge 15, Copenhagen 2450, Denmark
[20] Global Energy Systems Analysis, Technische Universität Berlin, Straße des 17. Juni 135, Berlin 10623, Germany
[21] Faculty of Economics and Social Sciences, University of Potsdam, August-Bebel-Str. 89, Potsdam 14482, Germany

E-mail: mathijs.harmsen@pbl.nl

Keywords: diagnostics, integrated assessment models, climate policy, 6th Assessment Report IPCC, renewable energy, mitigation, AR6

Supplementary material for this article is available online

Abstract

Integrated assessment models (IAMs) form a prime tool in informing about climate mitigation strategies. Diagnostic indicators that allow comparison across these models can help describe and explain differences in model projections. This increases transparency and comparability. Earlier, the IAM community has developed an approach to diagnose models (Kriegler (2015 *Technol. Forecast. Soc. Change* **90** 45–61)). Here we build on this, by proposing a selected set of well-defined indicators as a community standard, to systematically and routinely assess IAM behaviour, similar to metrics used for other modeling communities such as climate models. These indicators are the relative abatement index, emission reduction type index, inertia timescale, fossil fuel reduction, transformation index and cost per abatement value. We apply the approach to 17 IAMs, assessing both older as well as their latest versions, as applied in the IPCC 6th Assessment Report.

The study shows that the approach can be easily applied and used to indentify key differences between models and model versions. Moreover, we demonstrate that this comparison helps to link model behavior to model characteristics and assumptions. We show that together, the set of six indicators can provide useful indication of the main traits of the model and can roughly indicate the general model behavior. The results also show that there is often a considerable spread across the models. Interestingly, the diagnostic values often change for different model versions, but there does not seem to be a distinct trend.

1. Introduction

Integrated assessment models (IAMs) are widely used for climate policy and climate change analysis (van Beek *et al* 2020). They offer the means to assess the linkages between long-term climate policy goals and near-term policy choices. They can also look into mitigation strategies taking into account cross-sectoral and, cross-regional and systems interactions (energy, land, economy, climate). As such, they form a key information source feeding into the climate change mitigation policy process, e.g. via IPCC Assessment Reports (ARs) (Halsnæs *et al* 2000, IPCC 2014). Within IAMs, a distinction can be made between cost-benefit IAMs (mostly highly stylized) and detailed process IAMs that are mostly used to explore different pathways to reach selected policy goals. The latter comprise a diverse group of models with different functional structures.

A thorough understanding of how IAM structure and assumptions affect IAM behavior is critically important for assessing IAM based policy analysis and advice. For both policy makers and researchers, it can provide insights into why results differ between models and link projections to policy-relevant model assumptions and structure. It is the goal of diagnostic tools to foster such understanding. In fact, such tools can serve key functions: (a) characterizing model behavior by use of stylized diagnostic experiments, and (b) relating model behavior patterns to model structure and input assumptions. We focus mostly on the first in this study, but aim to cover the second, where possible. A subsequent function, but beyond the limits of this study is to qualify the model behavior and assess models' policy applicability.

In other modeling disciplines, similar diagnostic tools have been developed. For instance, in climate research, diagnostic metrics have been applied to compare climate models and to evaluate their performance (Andrews *et al* 2012, Flato *et al* 2013, Eyring *et al* 2016). Such indicators, for instance, include climate sensitivity (indicating the temperature increase for a doubling of the CO_2 concentration) and the transient climate response (indicating warming over a more limited time period). These tools are not only used to regularly compare models and thus qualify their behavior, but even in validation experiments, leading to assessment of the quality of models for specific experiments and their evaluation over time.

Also the IAM community has undertaken several model diagnostic activities in the past (Gaskins and Weyant 1993, Weyant 2004, 2010, van Vuuren *et al* 2009, Wilkerson *et al* 2015) resulting in the most recent and comprehensive diagnostic assessment by Kriegler *et al* (2015). Here, we propose an updated and expanded set of widely applicable, key diagnostic indicators to be used as a community standard. We determined these by revisiting the approach by Kriegler *et al* (2015) and improving them in terms of precision, simplicity and completeness. In particular, we propose a novel, standardized approach to compare different model versions to assess and monitor model differences over time. The approach is analogous to the climate model diagnostics in the sense that they are based on stylized scenarios with exogenous assumptions. It has been tested on 17 IAMs and 32 model versions, as part of two EU model development projects, ADVANCE (www.fp7-advance.eu/) and NAVIGATE (https://navigate-h2020.eu/), thus providing coverage of all main process-based IAMs (and much higher than in preceding studies), including all latest model versions. Especially the latter is highly needed in light of the forthcoming AR6.

A standard set of diagnostics for the community has obvious advantages. It provides a tool to systematically and consistently assess model behavior in all future studies. Model diagnostic results can be part of model documentation that can be referenced and highlighted in papers. Future model-intercomparison projects could require participating models to regularly run the core set of diagnostics, to analyze model behavior of newly developed models or model versions. Ultimately, this will lead to greater transparency and comprehensibility of IAM applications, together with model documentation. It will also allow tracking the development of IAMs over time—and possibly, in the future, confronting the outcomes with empirical information or information from other science disciplines.

An important innovation of the present study is the introduction of two diagnostic indicators in addition to the ones established by Kriegler *et al* (2015), namely inertia timescale (IT) and fossil fuel reduction (FFR). IT provides a measure of the models' level of inertia in response to the introduction of climate policy, a crucial determining factor in deep mitigation projections. FFR highlights the models tendency to reduce fossil fuels as part of climate policy, a key element in model studies that examine the energy transition.

Here, we present the results for six key indicators, adding IT and FFR to the original set of indicators from Kriegler *et al* (2015); relative abatement index (RAI), carbon intensity over energy intensity (CoEI), transformation index (TI) and cost per abatement value (CAV). The indicators have been simplified to make them more suitable to be used as a community standard, namely with a focus on one strong mitigation case and one benchmark year, 30 years in the future (here 2050, but later in post-2020 assessments). The latter allows for comparability with future diagnostic assessments. To ensure precision in the diagnostic results, we define single, unique values to indicate model behavior.

In method section 2, we explain the study design and list the participating models. The results are split-up in subsections for each of the indicators and conclude with an overview table to classify all the participating models. In the section 4, we reflect on the research questions: *Can these indicators be easily used as diagnostic tools for IAMs, including their development over time? And what insights do these tools provide?*

2. Methods

2.1. Diagnostic experiments and indicators

The experiments described in this study form a small selection from a larger set of stylized, diagnostic scenarios that have originally been developed as part of the EU FP7 ADVANCE project (www.fp7-advance.eu/). These are: Base (a zero carbon tax, i.e. a no-climate policy baseline) and C80-gr5 (a run with an exponential carbon equivalent price growth of 5% per year starting in 2020 and a price level of 80 (2010)\$/tCO$_2$ eq. reached in 2040). C80-gr5 is used for each key indicator presented here. For two indicators (RAI and IT) extra scenarios were used, as will be explained in the next section. Note that the C80-gr5 scenario represents a 1.5–2 degree case in most models (see supplement S7 (available online at stacks.iop.org/ERL/16/054046/mmedia)), in line with the Paris agreement's climate ambitions. This makes it a highly relevant showcase for assessing model behavior in frequent deep mitigation scenarios. Preferably, model groups used SSP2, the middle-of-the-road socioeconomic projection baseline scenario (Riahi *et al* 2017) for all assumptions, including population and economic growth.

The indicators are originally chosen and adapted here based on criteria set by Kriegler *et al* (2015):

- Identification of heterogeneity in model responses
- Diagnosis of relevant features for climate policy analysis
- Applicability to diverse models
- Accessibility and ease of use

Here, we add the following criteria:

- Standardization and comparability between diagnostic studies
- Precision/quantifiability

Based on these criteria, we derive a set of six indictors that describe model responses to climate policy. These indicators go beyond the work of Kriegler *et al*, because we provide a standardized formulation—in each case leading to a single value that characterizes the model. We specify set rules (benchmark year, scenario used, socio-economic assumptions) to allow for comparability between studies in a quantitative way. The main focus is on the year 2050 as it is (a) policy relevant and (b) provides a reasonable indication of model behavior throughout the century. For future use of the indicators, we define all indicators based on C80-gr5, using the value 30 years after the introduction of the tax (here 2020). While the focus is on 2050, we also show the 2100 results in the supplement (S3) to assess if the 2100 numbers would lead to different conclusions.

Table 1 gives an overview of the key diagnostic indicators proposed and assessed in this study. Below, we shortly summarize the setup and rationale behind the indicators and particularly indicate differences with and additions to the Kriegler *et al* (2015) approach. The combination of the indicators, focuses on (a) the responsiveness of the model, (b) the type of mitigation, (c) the scale of the transformation of the energy system, and (d) mitigation costs as a function of the carbon price signal.

As in earlier diagnostic exercises, the indicators are based on global totals to assess the overall behavior related to global climate policy. A regional assessment would be possible in a follow-up study. All emission indicators are based on CO$_2$ energy and industrial process (E&I) emissions. This allows for all models to participate (the land-use system and non-CO$_2$ emissions are modeled by about half of the models). Moreover, CO$_2$ E&I makes out more than two thirds of all GHG emissions (Olivier and Peters 2020).

The RAI characterizes the emission reductions in a carbon tax scenario relative to the baseline. It can be considered the main indicator in the sense that it measures the overall response to a climate policy incentive and correlates with elements from the other indicators (demand and supply side emission reductions, transformation rate, FFRs and limited inertia). Hence, it can also be considered a 'mitigation sensitivity' indicator, analogous to the 'climate sensitivity' in climate models. In order to assess mitigation of the full suite of GHGs, we also provide a full Kyoto GHG analysis in the supplement (S4). In addition, an additional scenario (C30-gr5, with a two thirds lower tax) is used to visualize a stylized 'derived MAC curve' from the RAI, by connecting the projected relative abatement at ~0, 50 and 130 \$/tCO$_2$.

Table 1. Key diagnostic indicators. For further explanation, see main text.

Indicator	Equation	Difference with Krieger et al. 2015	Short description
Relative Abatement Index (RAI)	$RAI(t) = \frac{CO_2\ E\&I\ Base\ (t) -\ CO_2\ E\&I\ Pol\ (t)}{CO_2\ E\&I\ Base\ (t)}$	None, apart from a focus on a single year (2050) and single scenario (C80-gr5)	Shows the relative reduction compared to the baseline. Indicates the "mitigation sensitivity" (as analogy to climate sensitivity)
Emission Reduction Type index (ERT)	$Carbon\ Intensity\ (t) = \frac{CO_2\ E\&I\ (t)}{Final\ Energy\ (t)}$ $Energy\ Intensity\ (t) = \frac{Final\ Energy\ (t)}{GDP\ (t)}$ $CIred\ (t) = \frac{(CI\ Base\ scen\ (t) - CI\ Policy\ scen\ (t))}{CI\ Base\ (t)}$ $EIred\ (t) = \frac{(EI\ Base\ scen\ (t) - EI\ Policy\ scen\ (t))}{EI\ Base\ scen\ (t)}$ $ERT\ (t) = \frac{CIred\ (t)}{(CIred\ (t) + EIred\ (t))}$	The original indicator was: Carbon intensity over Energy intensity, which did not strongly reflect reductions in energy efficiency. In addition, the new indicator focuses on a single year (2050) and single scenario (C80-gr5)	Shows the share of the RAI that can be attributed to decarbonization/energy supply side measures. 1 – ERT indicates the share of the RAI that can be attributed to reduced energy demand.
Transformation Index (TI)	$TI\ (t) = \lvert S1\ (t) - S1\ (2020)\rvert + \lvert S2\ (t) - S2\ (2020)\rvert + \ldots \lvert Sn\ (t) - Sn\ (2020)\rvert$ S = share of energy source in primary energy system	None, apart from a focus on a single year (2050) and single scenario (C80-gr5)	An index from 0 to 2 that indicates the overall transformation of the primary energy system.
Fossil Fuel Reduction (FFR)	$\mathrm{FFR}\ = \frac{(\text{Prim En fossil (2020)} - \text{Prim En fossil (t)})}{\text{Prim En fossil (2020)}}$	N.a., is new	Shows the relative reduction (%) in primary fossil fuel production compared to the base year
Inertia Timescale (IT)	$\mathrm{IT}\ = \frac{\text{Cumulative emission diff. (2040} - \text{2100)}}{\text{Emission gap (2040)}}$	N.a., is new	IT measures path dependency in terms of the convergence timescale after alignment of carbon price levels. It represents the number of years for emissions in a price shock scenario (C0to80-gr5, with no carbon pricing until 2040) to converge to an early carbon price scenario (C80-gr5) with an equal post-2040 price profile.
Cost per Abatement Value (CAV)	$\mathrm{CAV\ (t)}\ = \frac{\text{Mitigation Costs (t)}}{(\text{GHG Reduction (t)}\ *\ \text{Carbon Price (t)})}$	The original indicator was based on the net present value of future policy costs. Here, we do not apply discounting but focus on one benchmark year (2050) and one scenario ((C80-gr5)	Dimensionless indicator that shows the ratio between the policy costs and marginal abatement costs (higher = higher mitigation cost).

The ERT indicates the share of supply side measures (e.g. renewable energy) in bringing down emissions. 1 minus ERT shows the share of the RAI that that can be attributed to reduced final energy demand. Values higher than 0.5 imply supply models (= most common), lower than 0.5 imply demand models. This indicator replaces the CoEI indicator from Kriegler *et al* 2015): CI (as a fraction of CI in the baseline) over energy intensity, which did not strongly reflect reductions in energy intensity (e.g. a model with no energy efficiency at all could still be classified as a demand focused model).

Two energy system transformation indicators have been assessed: FFR, which is new in this study and transformation index (TI, from Kriegler *et al* 2015). FFR is a simple, policy relevant indicator that shows the relative reduction of fossil energy compared to the base year (2020). The FFR indicator was added to the transformation analysis, since it represents a less abstract alternative to TI and relates directly to recent studies aimed at fossil fuel phase out and renewable integration (in in the result section, we also compare FFR to TI to understand what drives transitions in models). TI shows the extent of transformation in the energy system (2 = max, 0 = none). Note that in table 1, the shares of energy sources in primary energy system (S), are based on the following aggregated energy sources: fossil,

Table 2. Participating models, types and versions. Latest model version indicated in bold. For detailed model documentation see: www.iamcdocumentation.eu/(IAMC wiki). See supplement (S1) for an overview of all scenarios and submissions by the different models.

	Model type	Solution Method	Time horizon	Analyzed versions
AIM/CGE – AIM/Hub	Recursive-dynamic CGE	Simulation	2100	V2, **V2.2**
EPPA	Recursive-dynamic CGE	Optimization	2100	**6**
FARM	Recursive-dynamic CGE	Simulation	2100	3.1, **4**
GEM-E3	Recursive-dynamic CGE	Optimization	2050	V2, **V2020**
IMACLIM	Recursive-dynamic CGE	Simulation	2100	**V1.1**
GCAM	Recursive-dynamic PE	Simulation	2100	4.2_ADVANCE, **5.3_NAVIGATE**
IMAGE	Recursive-dynamic PE	Simulation	2100	3.0.1, 3.0.2, **3.2**
POLES	Recursive-dynamic PE	Simulation	2100	ADVANCE, **NAVIGATE**
PROMETHEUS	Recursive-dynamic PE	Simulation	2050	V1, **V1.1**
iPETS	Inter-temporal GE	Optimization	2100	**V1.5**
MESSAGE-GLOBIOM	Inter-temporal GE	Optimization	2100	MESSAGE V4, 1.0, **1.1**
REMIND(-MAgPIE)	Inter-temporal GE	Optimization	2100	1.6, 1.7, **2.1-4.2**
WITCH	Intertemporal GE	Optimization	2100	4.2.0, **5.0.0**
COFFEE	Inter-temporal PE	Optimization	2100	**1.1**
DNE21+	Inter-temporal PE	Optimization	2050	**V.14**
TIAM-Grantham	Inter-temporal PE	Optimization	2100	**v3.2**
TIAM-UCL	Inter-temporal PE	Optimization	2100	3.1.5, 4.1.0, **4.1.1**

non-bioenergy renewables, bioenergy, nuclear, since these are reported by all models, thus allowing for a complete comparison.

In this study, we adopt a new indicator that describes the level of inertia (i.e. persistence of path dependency) in the models: IT. Path dependencies are of particularly relevance for the energy system, due to long-lived capital stocks, technological learning, and other sources of inertia in the upscaling of new technologies, as well as behavioral inertia on the demand side. They are also highly policy-relevant in the context of delayed climate policy adoption and carbon lock-in, as analyzed in several scenario studies (Riahi *et al* 2015, Luderer *et al* 2018). We here introduce a new diagnostic indicator that captures inertia in response to the introduction of climate policy as a crucial characteristic of IAMs. It is based on a newly introduced diagnostic carbon price shock scenario to quantify model representation of inertia. In our scenario set, the shock scenario follows baseline developments with zero carbon prices until 2040, followed by an instantaneous carbon price of 80$/tCO$_2$ in 2040, as in the default scenario, with an exponentially growing carbon price thereafter. For the shock scenarios, models with perfect foresight were instructed to disable the anticipation of future carbon pricing. The difference between the shock scenario and the default scenario can be measured in terms of the 2040 'emissions gap'. After 2040, the shock scenarios and corresponding early pricing scenarios can be expected to converge, since they are subject to the same carbon prices. However, during a transition period, the shock scenarios will continue to have higher emission levels than the corresponding early pricing scenarios, due to the systems inertia. The IT (in units of years) is defined as the ratio between the cumulative emission difference between the two scenarios after 2040, and the 'emissions gap' in the model year prior to 2040. For more information and visualization see supplement (S2).

The CAV is a dimensionless measure of economic implications of emissions abatement at a certain carbon price. It shows the ratio between the policy costs and marginal abatement costs (MACs). For PE models, this can be seen as an indicator for the shape of the (implicit) MAC curve. The closer to 1 this indicator is, the more concave the MAC curve and the higher the projected policy costs. In other words, a low value indicates more mitigation potential at lower carbon prices. For GE models, macro-economic feedbacks are also factored in. Here, a value higher than 1 implies that these feedbacks are a dominant factor in the costs. We simplified the original indicator by looking at a benchmark year (2050) instead of discounting to a net present value. Note that for this indicator, we include all greenhouse gases represented by the models (this differs per model), since that corresponds with the model's projected policy costs. Reported policy cost metrics also differ per model type. We used consumption loss compared to the baseline for all GE models and area under the MAC for all PE models, except for PROMETHEUS and TIAM-Grantham where the additional total energy system costs were applied. Although the metrics differ, they are comparable in the sense that they (at least) factor in first-order economic expenditures, which make out a considerable part of the policy costs.

2.2. Models

In total 17 IAMs, of which 32 unique model versions have participated in the diagnostic exercise, see table 2. The models have been broadly grouped based on their typology. One dimension in this typology is the coverage of the economy. Partial equilibrium (PE) models describe parts of the economy (e.g. such as the energy or agriculture sector) in detail, while having exogenous assumptions for the rest of the economy. PE models typically calculate climate mitigation policy costs as first order sector costs, such as area under the MAC curve for reducing greenhouse gases. General equilibrium (GE) models represent the full economy with varying levels of detail in the representation of sectors. GEs typically express policy costs in terms of consumption losses or GDP losses. The second dimension in the typology is the level of foresight in the solution function (for reaching climate targets), which is either high ('inter-temporal optimization' (ITO))) or low/myopic ('recursive dynamic' (RD)). RD models do not attempt to optimize costs over time, but use another set of rules for this. Dynamic recursive computational GEs (CGEs, see table 2) are a subgroup of GEs that follow such a myopic approach. These have a more detailed representation of sectors than ITO-GEs and derive costs based on deviation from market equilibria in individual years. The classification in table 2 is applied in all the analyses in this study, to determine any correlations between model type and behavior.

3. Results

3.1. RAI

Figure 1 shows the RAI in 2050 for different price levels (essentially showing a stylized derived MAC curve per model, 1a) and the RAI per model in the default scenario (1b). Models generally show the same characteristic at different price levels in 1a. As a result the RAI (1b) can be considered as representative for model response. When considering latest model versions, high RAI models (i.e. one standard deviation from mean) are IMAGE, REMIND-MAgPIE and AIM/Hub and low RAI models are POLES, IMACLIM and TIAM-Grantham. The high-low order in models generally persists at higher prices in 2100 (supplement S3) and when considering all greenhouse gases (supplement S4), implying that the 2050 CO_2-based benchmark is a relatively robust indicator. There is some indication that GE-ITO and PE-RD models have a relatively high response, while GE-RD models are generally lower—but there are large variations in each group.

There are considerable differences in RAI between model versions (notably of GEM-E3, MESSAGE, FARM and POLES) that can be traced back to specific model developments. However, there seems to be no consistent trend across the models towards either higher or lower abatement in newer model versions. The higher emission reduction achieved in the latest version of GEM-E3 is a result of improvements in representation of the energy system, especially in transport and in power generation. The new model version also captures the recent cost reduction of low-carbon technologies (e.g. photovoltaics (PV), wind, electric vehicles) thus enabling accelerated diffusion of these options. The lower abatement in the latest MESSAGE-GLOBIOM version results from model calibration (lowering the baseline emissions), reduced sustainable bioenergy potential and more pessimistic techno-economic assumptions on carbon capture and storage (CCS) deployment, despite more optimistic assumptions on non-bio renewables. Higher abatement in FARM results from more favorable CCS assumptions, both for fossil-electricity and bio-electricity. Lower abatement in the most recent POLES version is predominantly caused by slower deployment of CCS in power, industry and energy transformation (hydrogen, biofuels production). This outweighs several developments that increased abatement potential (inclusion of direct air capture, e-fuels and a more detailed representation of mitigation potential in final demand sectors (buildings, aviation, maritime and road transport). The low abatement potential in IMACLIM results from a persistence in fossil fuel use (see section 3.3).

3.2. Emission reduction strategy

Figure 2 shows the ERT in 2050 (2b) and underlying reductions in carbon intensity (CI) and energy intensity (EI) in the default scenario (2a). Note that all models can be considered supply models, i.e. that emission reductions are realized more via changes in energy supply (e.g. renewable energy) than in energy demand. This is indicated by all models being located right from the $x = y$ line in 2a and the >0.5 ERT values in 2b. Compared to the model mean, TIAM-UCL, GCAM, DNE21+ and IMACLIM can be considered high ERT models (strongly preferring supply side options) and POLES, MESSAGE-GLOBIOM and WITCH low ERT models (more demand-side focused). There is no apparent effect of model type on ERT.

At higher carbon prices (in 2100), supply side mitigation becomes more dominant for all models, indicated in 2a by the strong reduction in CI in 2100 compared to 2050, and higher ERT values in 2100 (see supplement S3). In IMAGE, higher prices also invoke a strong demand response, which is smaller for other models. In the case of REMIND, high prices even lead to an increase in energy intensity, caused by an increased energy demand for direct air capture and storage of CO_2 (DACCS, included in the last two

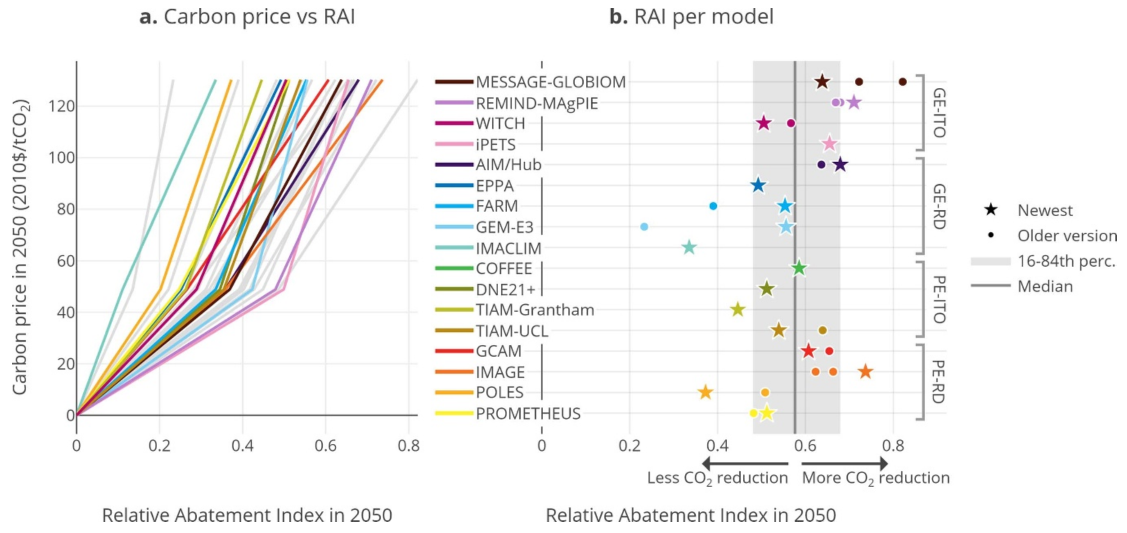

Figure 1. RAI in 2050. Left: RAI in 2050 for ∼50 and 130 \$/tCO$_2$. Coloured lines represent latest model versions. Right: RAI value for 130 \$/tCO2 based on default model run. Stars represent the latest model versions. Model are grouped based on type (see section 2). See supplement (S4) for the RAI based on the full suite of Kyoto GHGs.

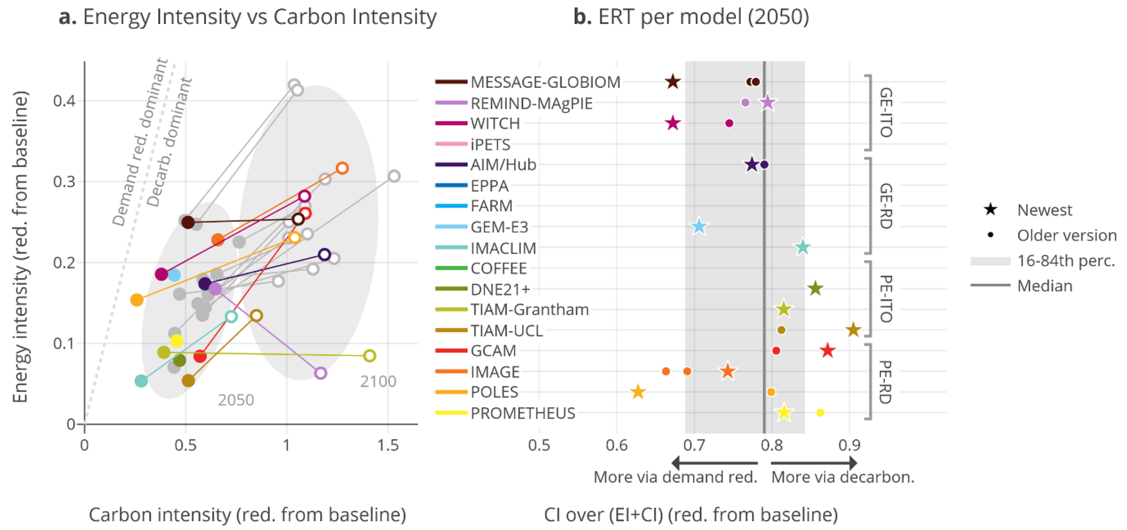

Figure 2. EI reduction vs CI reduction in 2050 and 2100 in the default scenario (left panel). Coloured lines represent latest model versions. Areas are 2-sigma confidence ellipses. Right: ERT in default run. Stars represent the latest model versions. Model are grouped based on type (see section 2).

versions) and to a lesser extent by higher electricity demand. The effect is magnified due to the exponentially growing carbon price and is less common in less extreme REMIND projections.

POLES and MESSAGE-GLOBIOM both show a large decrease in ERT compared to earlier versions. For both models, this is mainly caused by a decrease in supply side mitigation options (see description for RAI indicator). However, both models have a larger demand response in the latest versions.

3.3. Energy system transformation

The energy system transformation assessment in this study is based on two indicators: fossil fuel reduction (FFR, see figure 3(b)) and transformation index (TI, see supplement S5). Here we describe FFR and indicate large differences with TI, which signify a different level of transformation in the non-fossil parts of the primary energy system (renewables, bioenergy, nuclear). There is a large spread in FFR, varying from 43% reduction to an increase of 23%. Figure 3(a) (primary energy decomposition) shows that for most models, a considerable share of the remaining fossil energy consists of fossil energy without CCS. GE-ITO models generally seem to favor relatively high FFR. High FFR-models are REMIND-MAgPIE, iPETS, MESSAGE-GLOBIOM, TIAM-UCL, TIAM_Grantham and WITCH, Low-FFR models are IMACLIM, DNE21+, COFFEE and FARM. There is a high correlation between FFR and TI, as would be expected. A notable exception is COFFEE, which has a medium TI, due to large increases in bioenergy and to a lesser extent non-bio renewables. The high-low order of FFR and TI in models is very

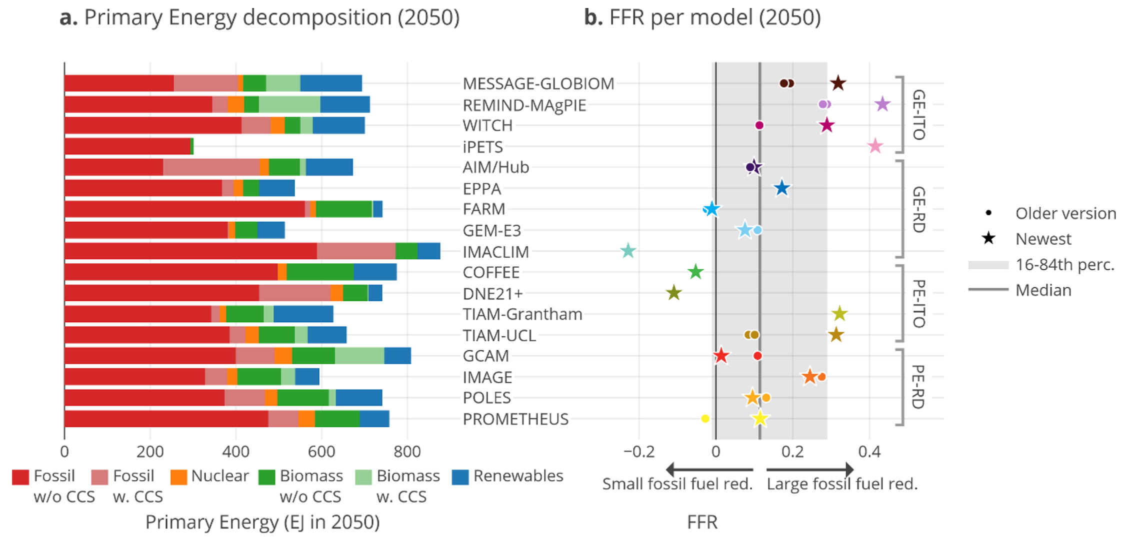

Figure 3. FFR indicator. Left: primary energy composition in 2050 in default run (only newest model versions). Note: where 'with CCS data' is missing, total fossil or bioenergy represents a combination of both with and without CCS. Note: iPETS nuclear and RE not shown (= 127 EJ in total). Right: relative FFR. Stars represent the latest model versions. Model are grouped based on type (see section 2). See underlying data, including the TI indicator in supplement (S5).

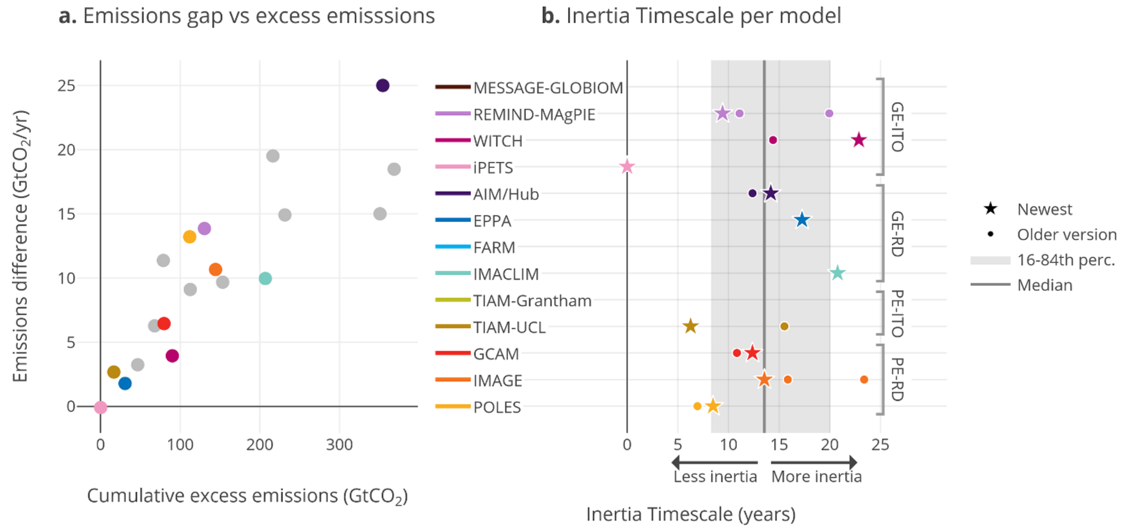

Figure 4. IT indicator. Right panel: IT per model. Stars represent the latest model versions. Model are grouped based on type (see section 2). Left panel: underlying components used to determine IT. *X*-axis: 'Additional emissions due to inertia' (i.e. cumulative CO_2 E&I difference between C80-gr5 and C0to80-gr5 in 2040–2100 period). *Y*-axis: difference in emissions between C80-gr5 and C0to80-gr5 in 2040. Colored dots represent latest model versions.

similar at higher prices in 2100 (see supplement S3 and S5) implying that the 2050 benchmark is robust.

Several large differences in model versions can be explained by model developments. The recent version of MESSAGE-GLOBIOM has more pessimistic assumptions about CCS, leading to stronger carbon price induced reduction in fossil fuel consumption. The increase of FFR in TIAM-UCL is caused by a reduction in capital expenditure for solar and wind and reduced growth constraints for renewables and CCS. High FFR in REMIND-MAgPIE is largely the result of a strong natural gas phase out and high renewables integration in the power sector. Similarly, in WITCH, it is due to updated renewables learning rates, and more costly assumptions about CCS storage. In contrast, COFFEE projects a large increase of natural gas, implemented as a mitigation option in industry sector, leading to a small increase of fossil fuels (= negative FFR). In DNE21+, fossil fuel persists due to an increase in oil demand and cost-effective mitigation via gas power generation, including CCS. Fossil fuel persistence is largest in IMACLIM, due to large capital inertia and myopic expectations of future carbon prices.

3.4. Inertia

Figure 4 shows the IT indicator for the models that took part in the inertia experiment (with 4b = IT and

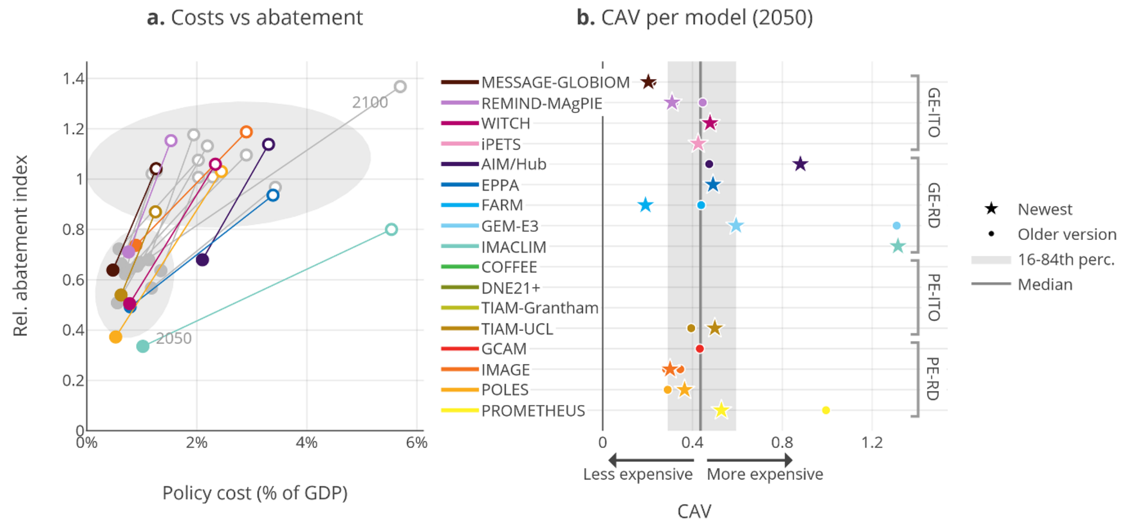

Figure 5. Policy costs vs relative abatement in 2050 and 2100 in default scenario (left panel). Coloured lines represent latest model versions. 2050 and 2100 areas are 2-sigma confidence ellipses. Relative cost (CAV) indicator by model in 2050 (right panel). Stars represent the latest model versions. Model are grouped based on type (see section 2).

4a = underlying data). Note that models with a 2050 time horizon are excluded, since IT is based on a full-century integral. There are large differences across models, with most showing IT in the 10–20 year range. iPETS is an extreme low-inertia scenario, with instantaneous convergence of the price shock and default scenario. TIAM-UCL and POLES also show a relatively low IT. WITCH and IMACLIM indicate the highest inertia. There is no apparent effect on model type on IT. Note that a sensitivity analysis in the supplement (S6) shows that the results are largely similar when the IT is based on a lower carbon price, implying that the default approach yields robust results.

Several large differences between model versions are the result of model developments. The more recent REMIND-MAgPIE versions favor electricity production from renewables, making it easier to reduce emissions in a short timeframe. Similarly, TIAM-UCL has reduced growth constraints for renewables and CCS, leading to a strong decrease in the IT. For the WITCH model, the current version has seen several updates based on latest insights: update of CCS storage potential (leading to less reliance on CCS in the second half of the century), renewables learning rates, short-term fossil fuel demand for India and China, introduction of time-varying elasticities of substitution and a reduction of the social discount rate to 2%–3%. This results in more stickiness of investments in the short term.

3.5. Policy costs

In figure 5, the policy CAV indicator is shown (5b) and a plot to visualize the policy costs (in % of GDP) versus the relative abatement in 2050 and 2100 (5a). The 2050 CAV is relatively comparable for most models, being in the 0.3–0.5 range, implying that the projected policy costs are around 30%–50% of the marginal costs. There is no clear trend towards either more costly or less costly mitigation in recent model versions. By design, GE models can produce higher CAV values, due to inclusion of macro-economic feedbacks. High CAV-models are IMACLIM and to a lesser extent AIM-Hub. In the case of IMACLIM, this results from assumed market imperfections in combination with imperfect foresight, leading to substantial GDP losses in a mitigation scenario. Notable low-CAV models are FARM and MESSAGE-GLOBIOM. The low-high model order in 2050 is almost identical to the order in 2100 (supplement S3), implying that the 2050 CAV provides a robust representation of the mitigation costs, also at high prices. Note however that the actual projected costs in a budget scenario (e.g. a 2 degree scenario) also depends on the assumed mitigation potential and carbon price.

Large CAV differences between model versions can be explained by the following model developments. The considerably lower CAV in GEM-E3 is mainly due to capturing of recent trends of cost reduction of low-carbon technologies (e.g. PV, wind, EV, batteries) as well as recalibration, which captures new trends of lower energy and carbon intensities. In the latest version of AIM/Hub, which represents an exception from the similar 2050–2100 behavior (showing a strong relative decrease in CAV in 2100), policy costs in 2050 are projected to be relatively high compared to 2100 due to limited availability of CCS (main factor) and bioenergy (the latter due to high population and lower yields).

3.6. Overview

Table 3 summarizes the classification of the models. Each model is indicated by a specific combination of six values that highlight its general responsiveness, the type of response and the responsiveness of the overall costs indicator.

Table 3. Overview of indicators & classification. All indicators are based on 2050 results (exception IT). 2100 results are shown in the supplement. Indicator acronyms from left to right: RAI, emission reduction type, FFR, TI, IT, CAV. Models are clustered based on type (general or partial equilibrium, recursive dynamic or intertemporal solution approach). Classification can be read as: (1) response based on RAI (2) emission reductions relatively high via energy demand (SD), supply (S), relatively strong supply (S+) based on ERT (3) policy CAV. High or low in the classification implies more than one standard deviation from mean. Grey is no data. Green/yellow highlight indicates: higher/lower value in a newer model version.

	Version	RAI	ERT	FFR	TI	IT	CAV	Classification
GE-RD models								
AIM/Hub	V2	**0.64**	0.79	0.09	0.33	11.5	0.47	Med response-S-Med CAV
	V2.2	**0.68**	0.77	0.10	0.35	14.2	0.88	Med response-S-High CAV
EPPA	6	**0.49**		0.17	0.26	17.3	0.49	Med response-S-Med CAV
FARM	3.1	**0.39**		-0.02	0.19			Low response-S
	4	**0.55**		-0.01	0.16		0.19	Med response-S-Low CAV
GEM-E3	V2	**0.23**		0.11	0.16	*		Low response
	V2020	**0.56**	0.71	0.08	0.51	*	0.59	Med response-S-Med CAV
IMACLIM	V1.1	**0.34**	0.84	-0.23	0.11	20.8	1.32	Low response-S+-High CAV
PE-RD models								
GCAM	4.2	**0.65**	0.81	0.11	0.47	10.8	0.43	Med response-S-Med CAV
	5.3	**0.61**	0.87	0.01	0.38	12.4		Med response-S+
IMAGE	3.0.1	**0.62**	0.66	0.24	0.40	23.4	0.35	Med response-DS-Med CAV
	3.0.2	**0.66**	0.69	0.28	0.47	15.3	0.28	Med response-DS-Med CAV
	3.2	**0.74**	0.74	0.24	0.40	13.2	0.30	High response-S-Med CAV
POLES	ADVANCE	**0.51**	0.80	0.13	0.40	6.4	0.29	Med response-S-Med CAV
	NAVIGATE	**0.37**	0.63	0.10	0.33	8.1	0.36	Low response-DS-Med CAV
PROMETHEUS	V1	**0.48**	0.86	-0.03	0.24	*	0.99	Med response-S+-High CAV
	V1.1	**0.51**	0.82	0.12	0.35	*	0.53	Med response-S-Med CAV
GE-ITO models								
iPETS	V.1.5	**0.66**		0.42		0.0	0.43	Med response
MESSAGE-GLOBIOM	V.4	**0.82**	0.77	0.19	0.56			High response-S
	1	**0.72**	0.78	0.18	0.39		0.22	High response-S-Med CAV
	1.1	**0.64**	0.67	0.32	0.52		0.20	Med response-DS-Low CAV
REMIND-MAgPIE	1.6	**0.68**	0.79	0.29	0.60	19.9	0.45	Med response-S-Med CAV
	1.7	**0.67**	0.77	0.28	0.53	10.5	0.32	Med response-S-Med CAV
	2.1-4.2	**0.71**	0.79	0.43	0.74	8.9		High response-S
WITCH	4.2.0	**0.57**	0.74	0.11		14.4	0.49	Med response-S-Med CAV
	5.0.0	**0.51**	0.67	0.29	0.46	22.9	0.48	Med response-DS-Med CAV
PE-ITO models								
COFFEE	1.1	**0.59**		-0.05	0.41			Med response
DNE21+	V.14	**0.51**	0.86	-0.11	0.09	*		Med response-S+
TIAM_Grantham	v3.2	**0.45**		0.32	0.52			Med response
TIAM-UCL	3.1.5	**0.54**		0.09	0.41			Med response
	4.1.0	**0.64**	0.81	0.10	0.43	15.5	0.39	Med response-S-Med CAV
	4.1.1	**0.54**	0.90	0.31	0.69	6.2	0.50	Med response-S+

* Cannot be measured because of the 2050 time horizon

4. Discussion & conclusions

Stylized diagnostic runs prove to be a useful tool to classify models (as in earlier studies) and to monitor model evolution, as we have shown here. There is a high demand for an approach to systematically and routinely assess model behavior in a standardized way. The method proposed here, with a focus on one benchmark year and standard scenario, allows for comparability between diagnostic studies over time with quantitative metrics. This study shows that the present +30 years benchmark provides a robust representation of model behavior over the century. We further showed that comparing different model versions based on the same experimental setup helps to understand model behavior, since changes can be traced back to specific model developments.

The focus here has been on the key indicators. However, the approach can be extended to secondary indicators that could provide sectoral or regional diagnostics and non-CO_2 greenhouse gases. Next to providing quantitative estimates of different aspects of model behavior, several key general conclusions can be drawn from this study's results:

- There is a considerable spread in outcomes for all indicators. This implies that the choice of a model in a study matters and that it is crucial to understand these differences.
- There is, however, no direct relationship between model type and model behavior (with some exception for GE models with intertemporal optimization that seem slightly more responsive).
- There does not seem to be a distinct trend in how models change in time with respect to the analyzed key indicators.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007-2013 under Grant Agreement No. 308329 (ADVANCE) and from the European Commission Horizon 2020 Programme H2020/2019-2023 under Grant Agreement No. 821124 (NAVIGATE). The findings and conclusions in this publication are those of the author and should not be construed to represent any official USDA or U.S. Government determination or policy. This research was supported by the U.S. Department of Agriculture, Economic Research Service. Shinichiro Fujimori is supported by the Environment Research and Technology Development Fund (JPMEERF20202002, JPMEERF20211001) of the Environmental Restoration and Conservation Agency of Japan.

ORCID iDs

Mathijs Harmsen https://orcid.org/0000-0001-6755-1569
Elmar Kriegler https://orcid.org/0000-0002-3307-2647
Detlef P van Vuuren https://orcid.org/0000-0003-0398-2831
Kaj-Ivar van der Wijst https://orcid.org/0000-0002-9588-7059
Gunnar Luderer https://orcid.org/0000-0002-9057-6155
Ryna Cui https://orcid.org/0000-0002-1186-8230
Laurent Drouet https://orcid.org/0000-0002-4087-7662
Johannes Emmerling https://orcid.org/0000-0003-0916-9913
Jennifer Faye Morris https://orcid.org/0000-0001-7675-558X
Florian Fosse https://orcid.org/0000-0002-0239-1143
Kostas Fragkiadakis https://orcid.org/0000-0002-1129-0360
Panagiotis Fragkos https://orcid.org/0000-0003-3596-0661
Oliver Fricko https://orcid.org/0000-0002-6835-9883
Shinichiro Fujimori https://orcid.org/0000-0001-7897-1796
David Gernaat https://orcid.org/0000-0003-4994-1453
Céline Guivarch https://orcid.org/0000-0002-9405-256X
Gokul Iyer https://orcid.org/0000-0002-3565-7526
Ilkka Keppo https://orcid.org/0000-0003-3109-1243
Kimon Keramidas https://orcid.org/0000-0003-3231-5982
Alexandre Köberle https://orcid.org/0000-0003-0328-4750
Peter Kolp https://orcid.org/0000-0003-0122-2839
Volker Krey https://orcid.org/0000-0003-0307-3515
Florian Leblanc https://orcid.org/0000-0001-9154-5847
Shivika Mittal https://orcid.org/0000-0003-4718-0064
Sergey Paltsev https://orcid.org/0000-0003-3287-0732
Pedro Rochedo https://orcid.org/0000-0001-5151-0893
Bas J van Ruijven https://orcid.org/0000-0003-1232-5892
Ronald D Sands https://orcid.org/0000-0002-2864-0339
Fuminori Sano https://orcid.org/0000-0002-7758-4441
Jessica Strefler https://orcid.org/0000-0002-5279-4629
Eveline Vasquez Arroyo https://orcid.org/0000-0002-2307-9757
Behnam Zakeri https://orcid.org/0000-0001-9647-2878

References

Andrews T *et al* 2012 Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models *Geophys. Res. Lett.* **39**

Eyring V, Bony S, Meehl G A, Senior C A, Stevens B, Stouffer R J and Taylor K E 2016 Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization *Geosci. Model Dev.* **9** 1937–58

Flato G *et al* 2013 Evaluation of climate models *Climate Change 2013: The Physical Science Basis Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* ed 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 (Cambridge: Cambridge University Press) pp 741–882

Gaskins D W and Weyant J P 1993 Model comparisons of the costs of reducing CO_2 emissions *Am. Econ. Rev.* **82** 318–23

Halsnæs K *et al* 2000 Summary report of IPCC expert meeting on stabilisation and mitigation scenarios: copenhagen *(Denmark, 2–4 June 1999)* (Roskilde: UNEP Collaborating Centre on Energy and Environment)

IPCC 2014 *Working Group III Contribution to the IPCC 5th Assessment Report "Climate Change 2014: Mitigation of Climate Change"* (Cambridge: Cambridge University Press)

Kriegler E *et al* 2015 Diagnostic indicators for integrated assessment models of climate policy *Technol. Forecast. Soc. Change* **90** 45–61

Luderer G *et al* 2018 Residual fossil CO_2 emissions in 1.5 °C–2 °C pathways *Nat. Clim. Change* **8** 626–33

Olivier J G J and Peters J A H W 2020 Trends in global CO_2 and total greenhouse gas emissions: 2020 report (The Hague: PBL Netherlands Environmental Assessment Agency)

Riahi K *et al* 2015 Locked into Copenhagen pledges—Implications of short-term emission targets for the cost and feasibility of long-term climate goals *Technol. Forecast. Soc. Change* **90** 8–23

Riahi K *et al* 2017 The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: an overview *Glob. Environ. Change* **42** 153–68

van Beek L, Hajer M, Pelzer P, van Vuuren D and Cassen C 2020 Anticipating futures through models: the rise of integrated assessment modelling in the climate science-policy interface since 1970 *Glob. Environ. Change* **65**

van Vuuren D P *et al* 2009 Comparison of top-down and bottom-up estimates of sectoral and regional greenhouse gas emission reduction potentials *Energy Policy* **37** 5125–39

Weyant J P 2004 EMF-19 alternative technology strategies for climate change policy *Energy Econ.* **26** 501–755

Weyant J 2010 Program on integrated assessment model development, diagnostics and inter-model comparison (PIAMDDI): an overview *IAMC Annual Meeting, 2010* (available at: www.globalchange.umd.edu/iamc_data/iamc2010/PIAMDDI_Overview.pdf)

Wilkerson J T, Leibowicz B D, Turner D D and Weyant J P 2015 Comparison of integrated assessment models: carbon price impacts on U.S. energy *Energy Policy* **76** 18–31