International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: repository@iiasa.ac.at
Web: www.iiasa.ac.at

Working Paper **WP-17-007**

Evaluating Process-Based Integrated Assessment Models of Climate Change Mitigation

Charlie Wilson
Elmar Kriegler
Detlef P. van Vuuren
Celine Guivarch
Dave Frame
Volker Krey
Timothy J. Osborn
Valeria Jana Schwanitz
Erica L. Thompson

Approved by

Arnulf Grubler
Acting Program Director, Transitions to New Technologies (TNT) Program

May 2017

Contents

Abstract

Process-based integrated assessment models (IAMs) analyze transformation pathways to mitigate climate change. Confidence in models is established by testing their structural assumptions and comparing their behavior against observations as well as other models. Climate model evaluation is concerted, and prominently reported in a dedicated chapter in the IPCC WG1 assessments. By comparison, evaluation of process-based IAMs tends to be less visible and more dispersed among modelling teams, with the exception of model inter-comparison projects. We contribute the first comprehensive analysis of process-based IAM evaluation, drawing on a wide range of examples across eight different evaluation methods testing both structural and behavioral validity. For each evaluation method, we compare its application to process-based IAMs with its application to climate models, noting similarities and differences, and seeking useful insights for strengthening the evaluation of process-based IAMs. We find that each evaluation method has distinctive strengths and limitations, as well as constraints on their application. We develop a systematic evaluation framework combining multiple methods that should be embedded within the development and use of process-based IAMs.

Keywords

integrated assessment models, evaluation

Acknowledgments

Charlie Wilson, Elmar Kriegler, Celine Guivarch, Volker Krey and Detlef van Vuuren acknowledge funding support from the European Union's Seventh Programme FP7/2007-2013 under grant agreement n° 308329 (ADVANCE). This paper builds on insights from participants at the Workshop on IAM Validation held in Seville, Spain in May 2013, particularly Yaman Barlas, Thomas Hertel, Anthony Jakeman, Robert Sargeant and Steve Smith. The workshop received funding support from the European Union's Seventh Programme FP7/2007-2013 under grant agreement n° 265139 (AMPERE).

Evaluating Process-Based Integrated Assessment Models of Climate Change Mitigation

Charlie Wilson [1 a,g], Elmar Kriegler [b], Detlef P van Vuuren [c,d], Celine Guivarch [e], Dave Frame [f], Volker Krey [g], Timothy J Osborn [h], Valeria Jana Schwanitz [i], Erica L Thompson [j]

1. Introduction

Evaluation of climate models using a range of methods is highly visible, organized in established programs, and synthesized in a dedicated chapter in each of the IPCC assessments of climate science over the past twenty five years (1). Climate model evaluation underpins model projections of long-term climate change (2).

The IPCC's assessments of climate change mitigation also draw heavily on modelling analysis, in this case by process-based integrated assessment models (IAMs) which characterize long-term transformation pathways in the energy and land-use systems (3). However there is no analogous synthesis of IAM evaluation in the IPCC assessments, nor elsewhere in the scientific literature. With the exception of model inter-comparison projects, IAM evaluation appears less systematic and less prominently reported.

We contribute the first comprehensive analysis of how process-based IAMs are evaluated, drawing on a wide range of examples across eight different evaluation methods testing both structural and behavioral validity. We use 'integrated assessment model' (IAM) to describe any model that: (1) explicitly represents the drivers and processes of change in global energy and land use systems linked to the broader economy; (2) captures both biophysical and socioeconomic processes including human preferences; (3) characterizes cost-effective mitigation pathways under different assumptions or constraints including climate stabilization targets (4). Over 1000 scenarios from 31 such IAMs form the basis of the IPCC assessment of transformation pathways to mitigate climate change (3, 5).

We follow the IPCC in making an important distinction between the process-based IAMs used in mitigation analysis which we consider in this article, and the highly-aggregated

[1] * corresponding author: charlie.wilson@uea.ac.uk
[a] Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, UK
[b] Potsdam Institute for Climate Impact Research, Germany
[c] PBL Netherlands Environmental Assessment Agency, the Netherlands
[d] Copernicus Institute of Sustainable Development, Utrecht University, the Netherlands
[e] Centre International de Recherche sur l'Environnement et le Developpement (CIRED), France
[f] New Zealand Climate Change Research Institute, Victoria University of Wellington, New Zealand
[g] International Institute for Applied Systems Analysis (IIASA), Austria
[h] Climatic Research Unit, School of Environmental Sciences, University of East Anglia, UK
[i] Sogn og Fjordane University College (HISF), Norway
[j] Centre for the Analysis of Time Series, London School of Economics and Political Science, UK

integrated assessment models used in a benefit-cost framework to analyze the economic impacts of climate change (6, 7). Confusingly both process-based and benefit-cost models are referred to in the literature as 'IAMs'. But as we explain in Box 1, the models used in a benefit-cost framework are very different tools which face a distinctive set of evaluation challenges (8). We do not consider these benefit-cost models further.

Process-based IAM analysis informs near-term decisions on energy and climate policy (9, 10), international negotiations on mitigation targets (11-13), and sustainable development strategies and goals (14-17).

These applications of modelling analysis depend on policy users' trust and confidence in IAMs as scientific tools that fulfil their intended functions (18). This confidence in the adequacy of IAMs is established and maintained by model evaluation involving both users as well as modelers.

Model evaluation begins with a clear statement of model purpose. IAMs are neither predictive nor directive. Rather IAMs serve as discursive tools to inform decision making on climate change mitigation (10, 19). IAMs sit alongside many other tools and approaches for informing climate policy, ranging from expert elicitations and bottom-up sectoral modelling, to learning from experience and participatory appraisals (20). However, IAMs are uniquely positioned to contribute insights on: (1) systemic effects, interactions, and trade-offs between sectors, regions, policy objectives and sustainability goals (21); (2) long-term consequences of near-term decisions (22); (3) process-based (causal) pathways to achieve predetermined global emission budgets (23).

We propose five inter-related criteria for assessing the adequacy of IAMs for providing policy-relevant insight. First, a model's application should follow logically from its purpose and design (*appropriateness*). Second, how a model conceptualizes and represents the modelled system should be clear in the analysis and communication of model output (*interpretability*). Third, model code should be clearly and transparently documented to enable independent review (*verifiability*). Fourth, users as well as modelers should have confidence in models as analytical tools good enough for their intended use (*credibility*). Fifth, models should advance understanding of policy options and challenges (*usefulness*).

The *credibility* and *usefulness* criteria are most closely linked to the application of uncertain model results in complex and contested policy domains like climate change (24, 25). Process-based IAMs have been critiqued for understating future uncertainties, for example, by forecasting future mitigation options and costs over a long (100 year) time frame (26). In this respect it is important to emphasize that evaluation does not make IAMs more accurate nor more reliable in predicting the future; this is not what IAMs are designed to do. Rather evaluation helps to improve the IAMs as scientific tools which are adequate for policy-relevant analysis.

As we argue in this article, a systematic evaluation framework helps IAMs improve against all five criteria: *appropriateness*, *interpretability*, *verifiability*, *credibility* and *usefulness*. To develop this framework, we consider the purpose, methods, and benefits of evaluating IAMs, drawing on a long tradition of IAM evaluation research (27). We first consider the role of structural and behavioral validity in IAM evaluation. We then review progress and best practice with eight distinct evaluation methods applied to IAMs, and identify challenges and limitations in each case.

Throughout the paper, we use climate model evaluation as a counterpoint to draw out particular issues with IAM evaluation methods. Climate models represent processes in the atmosphere and oceans (as well as biosphere and cryosphere) based on physical principles, and are used to simulate future climate change in response to natural and anthropogenic forcings (28). Both climate models and IAMs are tools used to advance scientific understanding of modelled systems and to provide policy-relevant insight. In the policy domain, both climate models and IAMs explore what-if questions given a set of assumptions (10, 29). Some IAMs include simple climate models. However IAMs are distinctive in modelling human preferences as well as natural processes. As we discuss, this and other differences have important implications for IAM evaluation compared with climate model evaluation.

Box 1. Benefit-cost integrated assessment models.

Highly-aggregated integrated assessment models applied in a benefit-cost framework are a distinct type of tool for informing climate policy. Benefit-cost models are used to estimate optimal mitigation efforts taking climate impacts on the economy into account (30).

Similar evaluation methods can be applied to both benefit-cost and process-based integrated assessment models. These methods include historical simulations, sensitivity analyses, and independent review (8, 31). However, the application and interpretation of these evaluation methods is fundamentally different between the two distinct types of model. As benefit-cost models lack a detailed representation of biophysical and socioeconomic processes in energy and land-use systems, they do not resolve the causal mechanisms driving greenhouse gas emissions in any detail. In contrast, it is precisely these causal mechanisms across different model components which are the emphasis of evaluation in process-based models.

Cost-benefit models also face specific modelling issues and challenges which further distinguish their evaluation needs (see SI for further discussion). In this article we only consider the evaluation of process-based integrated assessment models.

2. Evaluating Models of Complex, Dynamic Systems

Evaluation means assessing models and model performance so as to articulate the grounds on which a model can be declared good enough for its intended use (32).

Formally, evaluation tests the structural and behavioral validity of a model (33). Structural (or conceptual) validity means that a model is an accurate representation of the system response being modelled (34, 35) (see Box 2). Behavioral (or operational) validity means that modelled outcomes are consistent with observational data (see Box 3).

Behavioral validity is important for evaluating simulation models of well-defined, bounded systems with low uncertainties and the opportunity for reproducible experiments. These conditions do not apply to IAMs nor climate models which both represent complex, incompletely understood, dynamic systems. Consequently, structural validity is the stronger concept which precedes and to some extent subsumes behavioral validity (33).

However, the structural validity of complex models cannot be definitively established (see Box 2). Rather, evaluation is an open-ended process testing the adequacy of models' representation of the system *and* modelled system responses (33, 36). As Flato*, et al.* (1) argue: "*Climate models are based on physical principles and they reproduce many important aspects of observed climate. Both aspects contribute to our confidence in the models' suitability for their application ... for quantitative future predictions and projections*". The same ongoing process of evaluation applies to IAMs, particularly given the additional challenges they face in testing both structural and behavioral validity (see Boxes 2 & 3). The procedural nature of model evaluation necessarily involves a wide range of methods and activities in an iterative process of learning and improvement.

Box 2. Structural Validity of IAMs.

A model's structure or representation of the modelled system includes: variables; equations that encode laws, principles or causal relationships between variables; parameterizations that make simplifying assumptions about specific phenomena and include numerical parameters; and values assigned to input variables or parameters (37).

For models assessing complex environmental problems, including both IAMs and climate models, many of these structural elements are uncertain (25, 38). Epistemic uncertainties are associated with limits to knowledge of how the modelled system functions. Parametric uncertainties are associated with the reduction of complex phenomena to tractable model formulations and the values assigned to parameters (32). As a result of these uncertainties, structural validity cannot be definitively established. Models are not direct translations of scientific laws (39).

Compared to climate models, IAMs face an additional type of structural uncertainty. Socioeconomic processes are not based on physical principles or laws. The most appropriate representations of human preferences are changing and contested (40). Decision makers may even respond reflexively to modelling analysis, changing the relationships enshrined within the models. These are characteristic features of modelling in 'post-normal science' (41). Structural uncertainties are *societal* (related to social robustness and values embedded in model assumptions) as well as *epistemic* and *parametric* (related to ignorance and inexactness) (24). As IAMs cannot build on laws, theories, and principles to the same extent as climate models, establishing their structural validity is more problematic.

Box 3. Behavioral Validity of IAMs.

Behavioral validity is tested by tuning certain model inputs and parameters to match initial conditions and observed forcings over an evaluation period (42). The model is then run to test how well it predicts non-calibrated outcomes. Using historical simulations to test the behavioral validity of complex models such as IAMs or climate models has several limitations. First, simulation results may be specific to the tuned parameterizations and so reveal only 'forced empirical adequacy' (34, 43, 44). Second, more than one model conceptualization or parameterization can generate the same output, a problem known as 'non-uniqueness' or 'equifinality' (45, 46). This also implies difficulties in selecting from multiple possible combinations of parameters (47, 48). Third, two or more errors in the model inputs and parameterizations may partially cancel each other out (42, 49).

Consequently, a close fit of model output to observational data does not necessarily mean the model accurately represents the modelled system. The converse also holds. Divergence between model simulations and observations may be partly due to errors in inputs defining initial conditions or forcings. It is also not always clear whether and how to distinguish structural and parametric uncertainty as causes of divergence (47, 50).

All these issues apply generically to complex models including both IAMs and climate models. However, IAMs face additional issues in testing behavioral validity. First, historical simulations cannot demonstrate models' predictive reliability in future conditions that lie outside the range of historical experience (32). This is a particular issue for IAMs as the modelled system may not exhibit structural constancy between past and future (see Box 2). Human preferences expressed through climate policy may 'force' changes in the causal relationships enshrined in a model's representation of how the energy, land-use and economic systems function (51, 52). Second, IAMs are commonly used to define normative reference points such as least cost mitigation pathways. Normative applications of IAMs are not designed to reproduce observed system behavior. Third, IAMs may include optimization elements to capture price formation in markets. However, real markets are imperfect and IAMs may not capture the numerous distortions through which observed prices are reflected (53). Fourth, IAMs focus on system responses to policy forcings relative to a dynamic and uncertain baseline, rather than an equilibrium (54). As IAM baselines are dynamic, it is difficult to clearly separate model forcings (e.g., economic growth, prices) from system responses (e.g., energy resource use and technology deployment).

3. Evaluation of IAMs Compared to Climate Models

We compare evaluation methods applied to IAMs and climate models. We use this comparison to draw out distinctive challenges for evaluating IAMs. We first examine evaluation methods that use observational data, and then methods within or between models.

3.1. Historical simulations

Historical simulations are central to climate model evaluation. The recent IPCC assessment emphasizes the concurrent use of multiple indicators of model performance that are not related to tuned parameters and that span different processes, spatial scales and timescales (1). A common example compares observed anomalies in global mean surface temperature relative to a reference period against climate model predictions with anthropogenic and natural forcings over the period 1850 - 2005 (55). Simulated quantities like global mean surface temperature indicate trend responses to external forcings. Long-term simulations allow the response to climate forcings (the signal) to be more clearly separated from unforced internal climate variability (the noise). A close fit to observations builds confidence in a climate model's projections of the response to future emissions. Other climate model output, particularly over short timescales or small spatial scales, is related more strongly to internal variability. Examples include precipitation or El Niño events. Here, a close fit to observations builds confidence in a model's structural validity in representing coupled atmospheric-ocean processes.

Historical simulations with IAMs are less prominent than for climate models, and also tend to be more limited in time horizon, spatial scale, and model output compared to observations. Examples of simulated quantities in IAMs compared against observations include: energy use in US buildings during the period 1995-2010 (56); the Indian economy's response to rising oil prices during 2003-2006 (57); transportation energy demand in Western Europe during 1970-2003 (44). In each case, the simulations led to revised modelling assumptions to reduce divergence from observations (see SI for details and further examples).

However, the ability of IAMs to reproduce observations has weaker relevance as a test of behavioral validity for several reasons (see Box 3). These include the limited range of historical experience, the application of IAMs to define normative reference points, and the use of dynamic baselines. Historical simulations are therefore limited in their ability to give confidence in IAMs' predictive reliability under future conditions (58). As IAMs represent very diverse biophysical and socioeconomic processes as well as policy signals (59), simulated quantities must be sufficiently disaggregated to match this heterogeneity in underlying causal mechanisms (36). The causal mechanisms (or model component) should also be structurally constant over the simulation period. These limitations are compounded by significant data challenges for historical simulations of the energy, land use, and economic systems (56). Data challenges are more formidable in developing countries (60), and prior to the 1970s when few energy data were systematically collected (61).

Behavioral validity testing plays a less important role for establishing the structural validity of IAMs than it does for climate models. However historical simulations are an applicable evaluation method under certain conditions: (1) observational data are available; (2) forcings are clearly identifiable; (3) the structure of modelled system components is constant; (4) normative characteristics can be relaxed.

Climate model evaluation offers some useful insights for IAMs in developing historical simulations further. These include the importance of comprehensive long-term datasets of observations and forcings (59, 62), and standardized statistical measures of model performance against observations (e.g., correlations, root mean square percentage errors) to synthesize and visualize simulation results.

3.2. Near-term observations

The unfolding future provides near-term observations which can be compared against *ex ante* model projections made a decade or more ago. This is distinct from longer-term historical simulations which are run *ex post*.

Climate model projections from successive IPCC assessments have been compared against observed ranges at large scales across key outputs (e.g., mean global temperature and sea level rise) (Figs 1.4 & 1.10 in 63). Some climate models are also used for weather forecasting which provides numerous tests of very short-term predictions. However, internal variability limits inferences from near-term observations about model responses to forcings, especially at regional rather than global scales.

Baseline emission scenarios from the IPCC SRES were projected by IAMs in the late 1990s (64). These have similarly been tracked against actual socioeconomic developments and emissions since 2000 (65-67). Recent emission trends are towards the upper bound of *ex ante* projections across a range of baseline assumptions (68). One implication is that scenario studies may inadequately capture uncertainty ranges in key drivers or assumptions (69).

IAM projections of energy prices and demand have also been compared against observations (70, 71). However, these outcomes are characterized by short-term variability whereas IAMs are designed to represent long-term dynamics such as the replacement of capital stock and path dependence from increasing returns to scale (e.g., learning effects). Modelled responses to forcings in the near-term are not necessarily good indicators of long-term trends (59). Many IAMs also run on ten year time steps which capture only decadal averages.

Divergence from near-term observations is a potential source of insight for improving modelling efforts - if modelers look back at past projections (72). But this provides only a weak basis for testing the behavioral validity of IAMs. As with historical simulations, recent historical experience is useful for comparison against *ex ante* IAM projections only under certain conditions: (1) observations are linked to causal mechanisms with short-term characteristics and/or regional responses; (2) IAMs resolve processes in short time steps (1 - 5 years) or have structural elements responsive to short-term forcings; (3) the system response to policy forcings (e.g., renewable energy regulation) or exogenous shocks (e.g., oil crises, collapse of the Soviet Union) is clear and isolatable. The climate model community's recent experience in investigating and explaining the warming hiatus also shows the usefulness of open debate about causes of divergence between near-term observations and *ex ante* projections (73).

3.3. Generalizable historical patterns

An alternative method for drawing on history to evaluate IAMs examines whether generalizable historical patterns or 'stylized facts' are reproduced in model projections. This approach derives from the economist, Kaldor, who proposed "*a stylized view of the facts*" which held when observing economic growth over long time periods, ignoring business cycles or other causes of volatility (74, 75). An analogous evaluation method for climate models is subsumed within historical simulations which commonly include generalizable patterns as well as observations (76).

Schwanitz (36) proposed a set of generalizable historical patterns describing aggregate long-term behavioral features of the energy system and economy that are broadly applicable and expected to persist. Several studies have tested IAMs' ability to reproduce such patterns under both baseline and climate policy assumptions. Examples include: developing country transitions from traditional fuels to electrification as incomes rise (77); durations of technology diffusion correlating positively with extents of diffusion (78); primary energy consumption correlating positively with economic growth (36). In each case, model projections were broadly consistent with historical dynamics, albeit with local or spatial differences (see SI for details and further examples).

Rates of change in key system variables can also be compared between past and future to evaluate the responsiveness of actual and modelled systems. To-date, this method has been applied principally to IAM projections of technological change. Maximum projected rates of change are found to be broadly consistent with maximum rates observed historically, even in scenarios with stringent climate policy (79, 80).

These indirect tests of behavioral validity build confidence in IAMs' structural representation of long-term system dynamics. Generalizable historical patterns are an additional way for IAMs to draw on observational data while avoiding the limitations of historical simulations. But their application is also restricted to aggregate system-level indicators or relationships, rather than specific causal mechanisms. This exacerbates methodological difficulties in measuring and attributing divergence, and the implications this may have for structural validity.

3.4. Hierarchy of models (including simple models)

Confidence in a model's 'good-enough' representation of the modelled system can be effectively demonstrated through parsimonious models designed to capture only the fundamental drivers of change (81). Simpler models are also more amenable to independent review of underlying data and assumptions (70, 82).

Climate models span a range of complexity in terms of processes, dimensions, parameterizations, and spatial resolution (83). The resulting 'hierarchy of models' means the level of model complexity required for a given task can be defined by the research question (84). Climate models that are simpler, either conceptually, or in their resolution of processes and regions, remain useful to test understanding of the modelled system and so help interpret more complex models (43).

IAMs commonly have a reduced-form climate model component. Climate outcomes in IAMs across a range of emission scenarios have been tested against more complex climate models and found to correspond well (85). But climate model components represent only biophysical not socioeconomic processes.

There are few examples of IAMs stripped down to test the fundamental drivers of change in energy and land use systems. A notable exception is the 'SIMPLE' model of global agriculture which represents a minimal set of biophysical and economic relationships (86). A historical simulation from 1961-2006 was able to reproduce observed global trends in key indicators including crop land area, production, yield and price. Resulting confidence in the basic model conceptualization allowed critical assumptions embedded in more complex IAMs to be tested for their impact on model behavior (86).

This provides a powerful example of how highly simplified models can play an important role in testing structural validity within a hierarchy of models. IAMs largely lack such differentiated complexity. Drawing on climate modelling precedents, IAM evaluation should aim to: (1) develop reduced-form versions of complex models or incorporate simpler models from outside the IAM community to explore key structural assumptions; (2) match model complexity to the specific needs of a research question, trading off realism with interpretability (or what Held (84) describes as elaboration vs. elegance).

3.5. Model inter-comparisons

Model inter-comparison projects (MIPs) are used to explore structural uncertainties in different model representations of the same system. MIPs compare multiple models' outputs, insights, and fits to observations. To enable comparability, MIPs require carefully designed experiments that harmonize key scenario assumptions (including forcings) and standardize the reporting of model output (Section 1.5.3, 87, 88). Inter-model agreement can indicate results robust to structural uncertainty. However, agreement within the ensemble should be interpreted cautiously if structural differences between models are not systematic and models share approaches or components (50, 76).

MIPs are prominent evaluation methods for both IAMs and climate models, generating strong tacit learning for participating modelling teams. MIPs for climate models run in well-established programs (e.g., coupled MIP or 'CMIP'). Standardized CMIP output forms the backbone of the IPCC assessments of future climate change (2). As a recent example, CMIP5 organized ensemble runs of multiple models using prescribed sets of forcings and experiments, both historical and future-oriented (88).

Comparing results between multiple IAMs is a similarly longstanding feature of climate mitigation analysis (27, 89). IAM MIPs use controlled variations of policy assumptions (90, 91), technology assumptions (92, 93), or socioeconomic development assumptions (94) to explore ensemble uncertainties. Nine major IAM MIPs involving dozens of modelling teams contributed 95% of long-term mitigation pathways reviewed in the recent IPCC assessment (3, 5). Many of these focused on the robustness of policy-relevant insights such as the consequence of delayed global action on mitigation (95). As with climate MIPs, within-ensemble agreement in IAM MIPs is often interpreted as providing robust insights.

MIPs are the most well-established evaluation method for IAMs and are particularly useful for better understanding structural uncertainty. However methodological development is needed to help interpret inter-model agreement, taking into account structural similarities between models (e.g., shared components, similar representations of causal mechanisms). Climate model evaluation offers some precedent for interpreting results from model ensembles and attributing divergent results to structural differences between models (2, 96). Other techniques include aggregating or weighting IAM output within multi-model ensembles (97), or integrating expert judgements into the quantification of structural uncertainty (39).

An additional issue for IAM MIPs is how to address selection biases if stringent mitigation scenarios are not solvable by all models (98). In one IAM MIP exploring 2^{o}C stabilization, over a quarter of all model-scenario combinations were not solvable or not run (93). How to account for missing output data in IAM MIPs is a challenge sharpened

by the 1.5ºC climate stabilization ambition of the Paris Agreement and the extremely constrained emission budgets that this implies.

3.6. Diagnostic indicators

Model diagnostics are a specialized application of MIPs that use a standardized set of indicators or performance metrics. These indicators classify model behavior under harmonized forcings or scenario assumptions (99). Diagnostic indicators therefore serve to 'fingerprint' models.

Prescribed diagnostic runs are a precondition for climate models to participate in CMIP. Performance metrics span a range of model processes and functions. The most prominent example is climate sensitivity which is the mean global temperature change after doubling atmospheric CO_2 concentrations above pre-industrial levels. Cloud feedbacks represent the main cause of variation between models in this aggregate diagnostic indicator (100).

Analogous diagnostic runs in IAM MIPs use prescribed carbon price forcings under harmonized economic growth and demographic assumptions (101, 102). Diagnostic indicators include the aggregate economic cost of mitigation and the extent of transformation in the energy system. Although descriptive, these indicators are an enabling step towards explaining characteristic model performance in terms of model structure and assumptions. As an example, Wilkerson, Leibowicz, Turner and Weyant (102) attribute inter-model variation in emission outcomes to differences in the portfolio of available low-carbon energy supply options and the adaptability of electricity networks.

Diagnostic indicators offer a standardized and transparent complement to MIPs for clearly representing differences in model behavior. Model fingerprints also enable specific models to be selected to match the analytical needs of specific scientific or policy questions. Diagnostic indicators have only recently been developed for IAMs. Consolidation of community-wide standards would enable their integration within IAM evaluation efforts, and potentially their application as preconditions for participating in major IAM MIPs (as with CMIP). This should include a systematic characterization of the major elements of IAMs (including structure, parameterization, and input assumptions) so that explanatory links between model designs and model fingerprints can be mapped.

One restriction on the use of diagnostic indicators in IAM MIPs is that IAMs vary widely in design, so only a limited subset of forcings can be harmonized across a model ensemble. As an example, GDP growth is an exogenous forcing for some IAMs, but endogenously generated in others. This limits the extent to which diagnostic differences can be comprehensively explained.

3.7. Sensitivity analysis

Sensitivity analysis is used to: (1) identify the model inputs and parameterizations influential on model output; (2) attribute uncertainties in outputs to uncertainties in inputs. Local or ‘one-at-a-time’ methods test output sensitivities to changes in single inputs or parameters; global methods vary many or all inputs or parameters simultaneously (103).

The local sensitivity of climate model output to uncertain parameters is tested in perturbed physics ensembles (PPEs) in which key atmospheric, ocean or land-use parameters are

varied (2, 96). Exploring the full uncertain parameter space through repeated model runs is computationally costly, and prohibitively so for global methods. Distributed computing power offers a way round this constraint, as demonstrated by climatepredicton.net (104, 105). Alternatively, statistical models (emulators) can be fitted to the relationships between parameter values and model output from a smaller PPE allowing model sensitivities to be generalized to unexplored parameter values (50).

Local sensitivities in IAMs are tested as part of policy applications, but are more commonly reported in separate model evaluation studies. Influential inputs or parameters include rates of technological change (4), rates of energy efficiency improvement (106), and investment costs of energy supply technologies (92, 107, 108). However, local sensitivity analyses on discrete parameters provide limited insights for structural validity (109).

Global sensitivity analyses are also possible in IAMs using computationally-efficient techniques (110). These have recently been used to explore the multi-dimensional global space spanned by uncertain model inputs and parameters (92, 111) (see SI for details and further examples). This is a promising avenue of research in which IAMs have a relative advantage over climate models of greater computational complexity. Sensitivity analysis opens up the interpretability of IAM results in terms of input assumptions, particularly if reported alongside model applications (112). The potential for global sensitivity analysis in IAMs lessens the relevance of climate model precedents (e.g., distributed computing, emulators).

3.8. Model documentation, checks & review

A range of quality control procedures internal to modelling teams support model evaluation, particularly with respect to structural validity (35). These include checks of computer code and model implementation, as well as peer review or expert appraisal of how models conceptualize and represent the modelled system (38, 39).

More consistent and complete reporting of input assumptions, parameterizations, and model documentation and code also help open up models to independent third party review (51, 113). However, it is costly to make accessible complex and computationally-intensive models. Modelling teams may also seek to protect the intellectual property invested in model development (114). Open access models may allow repeatable model experiments but place a high burden on available resources to maintain the models and support users.

In general both climate models and IAMs face similar challenges for improving transparency. Publicly-available model documentation and databases of standardized model output were a central feature of CMIP5 (88). Databases of IAM output from several MIPs were similarly published online prior to the recent IPCC assessment. Many IAM teams have long made available extensive technical documentation (e.g., 115, 116). Consistent and comparable IAM documentation is also becoming more common, and published online in standardized wiki format (see SI for details). Some IAMs (e.g., GCAM) and climate models (e.g., CESM) do have publicly-accessible source code and data. Climate modelling teams are increasingly publishing details of how they tune parameterizations for processes like cloud formation or sea ice reflectivity (114).

Independent review is unique among the evaluation methods in supporting verifiability (51). This in turn builds confidence in the structural validity of IAMs among a potentially

diverse range of modelers, domain experts, and users. However in the IAM community, evaluation methods are more commonly used by the modelling teams and reported as part of model applications (51). This restricts the breadth of independent expertise involved in IAM evaluation.

4. A Systematic Approach for Strengthening IAM Evaluation

4.1. A systematic evaluation framework for IAMs

Each IAM evaluation method has certain restrictions on its application, and limitations in what can be learnt about structural or behavioral validity. That is not to say that any given evaluation method is inapplicable or irrelevant; rather that multiple evaluation methods should be applied in concert so that the limitations of one are addressed by the strengths of another. Table 1 synthesizes the strengths and limitations of the eight evaluation methods considered. The limitations define conditions under which specific methods should be applied.

The distribution of strengths (and limitations) across different evaluation methods makes it important to synthesize and compare insights on models and model performance. The prominent syntheses of climate model evaluation research in the IPCC assessment reports provides a useful precedent.

In practice therefore, a systematic approach to IAM evaluation consists of: (1) applying specific evaluation methods, subject to their specific limitations; (2) documenting learnt insights on model structure and function, and resulting model improvements; (3) synthesizing and comparing insights across methods (both within and between models); (4) involving user communities in interpreting and communicating evaluation insights; (5) identifying gaps or ongoing evaluation needs. This implementation process is iterative and continual (36). Evaluation establishes a direction of travel not a destination.

Table 1. Strengths & limitations of eight IAM evaluation methods.

Evaluation method	Strengths	Limitations
historical simulations	- use of observations - behavioral validity test	- not as relevant for normative model applications (e.g., cost-effective pathways) - not predictive of future conditions - difficulty of separating forcings from system responses in dynamic baselines - limited to specific causal mechanisms or spatial scales - limited implications for structural validity (and issues with over-tuning)
near-term observations	- use of observations - behavioral validity test	- models not designed to capture short-term variability - (*other limitations as for historical simulations*)
generalizable historical patterns	- use of patterns in observations - broad applicability (less sensitive to specific historical conditions and data constraints) - contributes to behavioral validity	- subjective comparisons, no standardized tests - difficulty of identifying reasons for divergence and implications for structural validity
hierarchy of models & simple models	- tests understanding of key system processes - links model behavior to structural validity - clearly interpretable results	- limited applicability and policy-relevance - difficulty in simplifying heterogeneous causal mechanisms
model inter-comparison projects (MIPs)	- identify results robust to (and sensitive to) structural uncertainty - insights on structural validity - peer review, exchange of data and methods, tacit learning among modelling teams	- limited standardization of scenario implementation across diverse models - difficulty of attributing divergent results to individual model differences - risk of groupthink in shared modelling strategies, removal of outliers
diagnostic indicators	- standardized and comparable model performance metrics - generalizable model classification or 'fingerprint' - link differences between models' structure and parameterization to differences in models' behavior - insights on structural validity	- descriptive indicators of model behavior, not explanatory - risk of over-tuning to harmonize diagnostic model runs - not appropriate or possible for all models with different designs
sensitivity analysis	- identifies influential inputs and assumptions - links model inputs and parameterization to model behavior	- does not address structural uncertainty in models - computational cost of global methods - limited insights from local methods
documentation, checks, review	- third party verification & expert review - transparency, openness	- costly (time, capacity, intellectual property)

4.2. Improving IAMs against five evaluation criteria

Five criteria provide the dimensions along which evaluation can help improve IAMs and their use in policy contexts: *appropriateness*, *interpretability*, *verifiability*, *credibility*, *usefulness*. Table 2 summarizes the contributions each evaluation method makes on each criteria. No single method addresses all five criteria.

Table 2. IAM evaluation criteria and methods.
Correspondence between criteria and methods is subjectively labelled as strong (green ☐), partial (amber ~), or weak/none (red) based on evidence presented in previous sections.

IAM evaluation criteria		using observations			within or between models				
IAM evaluation methods		historical simulations	near-term observations	generalizable historical patterns	hierarchy of models	model inter-comparisons	diagnostic indicators	sensitivity analysis	documentation, checks & review
appropriateness	is model purpose and design consistent with research question?	☐	~	☐	☐		☐		~
interpretability	are model results interpretable in light of model structure & parameters?	~		~	☐	☐	☐	☐	~
verifiability	are model results repeatable or model structure accessible to 3rd parties?				~				☐
credibility	is model good enough for its intended purpose for both users and modelers?	☐	~	☐	~	☐	~	~	☐
usefulness	do model insights advance understanding of policy options & challenges?		~		~	☐		☐	

Model evaluation should help provide a clear statement of a model's *appropriateness* for addressing a specific scientific question (35, 81). Ever more extensive, higher resolution representations of coupled natural-human systems creates IAMs with numerous purposes (117). Evaluation methods that delineate specific characteristics of models or model performance support *appropriateness* in matching tool with task. Diagnostic indicators help select IAMs with specific performance characteristics to answer related policy questions. A hierarchy of models allows simpler, more clearly interpretable IAMs to be used for characterizing general system dynamics. IAMs with specific causal mechanisms tested against observations in historical simulations are appropriate for policy analysis linked to those mechanisms.

Model evaluation should improve the *interpretability* of results, taking model structure and assumptions into account (118). Model inter-comparison projects and diagnostic indicators contribute to the *interpretability* of IAMs by linking model behavior and resulting policy insights to structural representations of energy, land use and economic processes. Sensitivity analysis similarly links model behavior to input assumptions and parameter values. As IAMs represent multiple systems and their interactions, another pragmatic approach for improving *interpretability* is to evaluate individual components sequentially such as their climate models (85, 119).

Model evaluation should improve the *verifiability* of model structure and results in line with scientific best practice (51). More transparent, documented and accessible models and model results support the *verifiability* of IAMs. Open source models further enable the repeatability of model experiments outside the parent modelling teams. Standardization also supports transparency and interpretability. Performance metrics could be developed and applied to standardize IAM evaluations, building on recent work with diagnostic indicators (101, 102).

The two remaining criteria, *credibility* and *usefulness*, are particularly important for policy-relevant models, and emphasize the importance of users as well as modelers in evaluation (35). How the producers and users of knowledge interact may be as important in determining the *credibility* of IAMs as the modelling analysis itself (120, 121). Modelers should clearly communicate "insights not numbers" (19) alongside assumptions, uncertainties and limitations (25, 122, 123). Boundary organizations at the science-policy interface play a critical translational role. Dedicated IPCC chapters are important signals of concerted evaluation efforts.

Finally, as Barlas and Carpenter (18) observe: "*Models are not true or false but lie on a continuum of usefulness*". Evaluation should underwrite the *usefulness* of IAM analysis. IAMs are designed to capture interdependencies, feedbacks, trade-offs, and the systemic consequences of climate policy at a global scale over the long-term (21). It follows that IAMs' *usefulness* lies primarily in applications linked to these unique design features. The future possibility space explored by IAMs is wide, with irreducible uncertainties in both biophysical and socioeconomic systems (25, 124). The *usefulness* of IAMs lies also in characterizing salient uncertainties, and identifying the robustness of alternative policy options for achieving a given climate stabilization goal (58, 125).

4.3. Applying insights from climate model evaluation

Climate model evaluation yields relevant methodological and practical experience which can inform IAM evaluation. First, the use of historical observations, whether through simulations, near-term observations or generalizable patterns, are important for behavioral validity testing despite the many restrictions on their use. The historical record remains the only basis for observing how modelled systems actually behave.

Second, both long-term observational datasets, and simple as well as complex models are useful for testing model representations of key system processes. Third, standardized performance metrics are useful for transparently comparing between models. Fourth, the involvement of independent research groups in evaluation research increases the visibility of model evaluation, as well as the independence and diversity of expert knowledge brought to bear on model development (1).

However, in drawing on experiences from climate model evaluation, it is important to recognize important differences between IAMs and climate models that help explain why certain evaluation methods are less prominently applied to IAMs.

One set of differences relate to the modelled system. The climate system is based on invariant physical principles which ensure structural constancy (between past and future). Long-term observations of global mean surface temperature provide a good basis for testing model response to forcings relative to a pre-industrial equilibrium. In contrast, IAMs represent socioeconomic processes which may not be structurally constant under climate policy, and baselines are dynamic and uncertain (see Box 2 and 3).

A second set of differences relate to domains of application. Climate models provide unique insights with few competing sources of knowledge (126). Understanding future climate change and impacts rests heavily on climate model projections (76). Users skeptical of climate science have sharpened the perceived need for model evaluation by climate scientists (114). In contrast, IAMs provide insights on mitigation options and challenges to decision makers with relevant own expertise as well as access to many different sources of knowledge and analytical tools (97).

A third set of differences relate to the modelling communities. The computational requirements of climate modelling and the availability of funding to support large centers means evaluation research is better supported. In contrast, the IAM community is more dispersed and funding is not typically available to support model evaluation (127).

The combined effect of these differences has been to mainstream and standardize evaluation methods within the norms, agendas and research activities of the climate modelling community to a greater extent than with IAMs.

5. Conclusion

Process-based IAMs are one of many analytical tools and sources of policy-relevant insight on climate change mitigation. However, IAMs are unique in being able to analyze systemic outcomes of coupled socioeconomic and biophysical processes in response to policy forcings. IAMs provide quantitative analysis of associated uncertainties to policymakers over the long time horizons of climate change mitigation, and characterize energy and land use transformation pathways consistent with predetermined emission budgets. It is important to continually test and improve the adequacy of these tools.

Evaluating IAMs helps establish the legitimacy of their use, the appropriateness and adequacy of their application, and confidence in their results among users. We have synthesized many examples, benefits, and limitations of applying different evaluation methods to IAMs. The time is now ripe for establishing a more systematic approach to IAM evaluation, combining different methods in an ongoing, collaborative process involving both modelers and users. Such a systematic approach is needed to improve IAMs' appropriateness, interpretability, verifiability, credibility, and usefulness.

6. Supplementary Information (SI)

(1) Benefit-cost integrated assessment models.

(2) Examples of historical simulation studies by process-based IAMs.

(3) Examples of generalizable historical patterns tested in process-based IAMs.

(4) Characteristics of process-based IAMs participating in two recent model diagnostic studies, including key references and links to model documentation.

(5) Examples of global sensitivity analyses by process-based IAMs.

(1) Benefit-cost integrated assessment models.

Highly-aggregated integrated assessment models applied in a benefit-cost framework are another type of widely-used tool to inform climate policy (30, 31, 127). Benefit-cost models are used to estimate optimal mitigation efforts taking climate impacts on the economy into account. Example of benefit-cost models include DICE (128) and PAGE (129). In contrast, process-based models are used to find least-cost pathways to achieve a pre-determined level of mitigation effort such as 2^{o}C climate stabilization. Examples of process-based models include IMAGE (17), REMIND (130) and GCAM (79). Kunreuther, *et al.* (20) discusses the applications and limitations of process-based and benefit-cost integrated assessment models in mitigation analysis (also see Stanton, Ackerman and Kartha (131) for a comparative review).

Although similar evaluation methods can be applied to both benefit-cost and process-based models, cost-benefit models face additional and specific modelling issues which distinguish their evaluation needs (132). These modelling issues include: (1) the sensitivity of results to discount rate assumptions (133); (2) the atheoretical and weakly empirical basis of 'damage functions' which parameterize the impacts of climate change on the economy (131); (3) the omission of tipping points and potential catastrophic impacts of climate change (134-136); (4) the weak integration of mitigation co-benefits in a welfare maximization framework (137). Discount rates and damage functions in particular strongly influence model estimates of the social cost of carbon which defines the economically-optimal level of mitigation effort (138, 139).

(2) Examples of historical simulation studies by process-based IAMs.

TABLE S1. IAM BEHAVIOR TESTED AGAINST OBSERVATIONS.

Study	Global IAM	Timescale & Spatial Scale	IAM Behavior & Variables Tested	Key Findings -> Resulting Learning Outcomes
(57)	IMACLIM-R (global)	2003-2006 India	Short-run macroeconomic response to rising oil prices as exogenous shock	Over-estimation of negative impact on economic growth -> modifications of model parameterization
(140)	GTAP-E (global)	1980-2005 Global	Distribution of petroleum prices in response to demand and supply shocks	Over-estimation of medium-run demand elasticities -> modifications of model parameterizations
(141)	TIMER/ IMAGE (global)	1970-2003 USA, W. Europe, Brazil, India, China	Energy demand in residential sector	Uncertainties in calibration parameters to reproduce historical data -> importance of reporting calibration ranges
(44)	TIMER/ IMAGE (global)	1970-2003 W. Europe & India	Energy demand for transportation	Input parameters calibrated to reproduce historical data -> but wide ranges of possible parameter values in multiple combinations
(56)	GCAM (global)	1995-2010 US	Energy demand in buildings	Accurate reproduction of growth in floor space, but over-estimation of electricity use and under-estimation of gas use -> modifications of model parameterization
(142)	AIM/CGE (global)	1981-2000 Global & Regional	Primary energy mix Electricity supply mix Final energy mix by energy sources and end-use sectors	Good reproduction of global totals but larger errors at the level of specific resources or technologies, and a disaggregated regional scale -> discussion of possible structural uncertainties (e.g., non-price factors in power plant allocations) as well as parametric uncertainties (e.g., oil price elasticity, autonomous energy efficiency improvement)
(47)	CIMS * (Canada)	1990-2010 Canada	Consumer choices of energy end-use technologies	Technology choice functions calibrated to reproduce historical data -> but convergence on a set of preferred probability distributions only for some technologies (heating systems) not others (refrigerators, cars)

* Not a global IAM but included for comparison purposes in key findings and resulting learning outcomes.

(3) Examples of generalizable historical patterns tested in process-based IAMs.

Kaldor's stylized facts included the rate of return on investment being roughly constant over long periods, and the real wage growing over time. For an updated set of Kaldor's stylized facts, see (143). Several of these generalizable historical patterns describing economic growth have been tested against regional GDP projections used to generate future emission scenarios (75).

Schwanitz (36) tests two generalizable historical patterns relating to energy technology and consumption using global integrated assessment models. The first is that variation in per capita growth rates increases with distance from the technology frontier (143). The second is that primary energy consumption is positively correlated with economic growth (144). Using model results from the Global Energy Assessment (14), Schwanitz (36) finds that the first generalizable historical pattern is not reproduced and the second is broadly reproduced other than during a 30 year period in a particular model region. In the first case, economic convergence is an exogenous input to the models, suggesting the usefulness of alternative per capita growth assumptions at a disaggregated regional scale. In the second case, further testing is needed to determine whether the model deviation from the historical pattern is plausible (36).

Wilson, Grubler, Bauer, Krey and Riahi (78) use historical data from a sample of 9 energy technologies to characterize a generalizable pattern of technological lifecycles: the extent to which an energy technology diffuses (measured in terms of cumulative capacity) is positively correlated with the duration of its diffusion. This pattern has more recently been confirmed with an expanded sample of 16 energy technologies (145). Using results from two IAMs under both baseline and mitigation scenarios, Wilson, Grubler, Bauer, Krey and Riahi (78) find that models tend to show longer durations of diffusion for a given extent of diffusion than those observed historically, although the effect of emission constraints is to compress these durations. This divergence in model behavior from a historical stylized fact can likely be attributed to model preferences for continuous, balanced and concurrent growth among technologies within a more diverse portfolio than has been observed historically.

(4) Characteristics of process-based IAMs participating in two recent model diagnostic studies, including key references and links to model documentation.

Tables S2 and S3 summarize key characteristics of global IAMs that participated in two recent model diagnostic studies (101, 102). This is a limited sample of global IAMs included here for illustrative purposes. For a more complete synthesis, see (5).

It is important to note that the system boundaries of many global IAMs can vary. As an example, models of the energy system and economy (e.g., MESSAGE, REMIND) can be coupled to land use models to endogenize land use dynamics (e.g., MESSAGE-GLOBIOM, REMIND-MAgPIE). Tables S2 and S3 report only the characteristics of models used in the diagnostic studies.

Further details about the models, including their different configurations, can be found in the detailed model documentation reported below the tables.

TABLE S2. GLOBAL IAMS PARTICIPATING IN KRIEGLER, *ET AL.* (101) MODEL DIAGNOSTIC STUDY. SOURCE: (101, 146). NOTE: MAC = MARGINAL ABATEMENT COSTS.

Global IAM	Modelling team home institute	Equil-ibrium type	Modelling approach	Time hor-izon	Resolution of energy supply	Representation of land use *	Coverage of greenhouse gases
AIM-Enduse	NIES, Japan	Partial	Recursive dynamic	2050	High	None	Kyoto gases
DNE21+	RITE, Japan	Partial	Intertemporal optimization	2050	High	MACs for land use emissions	All GHGs and other radiative agents
GCAM	PNNL, USA	Partial	Recursive dynamic	2100	High	Endogenous land use dynamics, afforestation	All GHGs and other radiative agents
GEM-E3	ICCS, Greece	General	Recursive dynamic	2050	Low	MACs for land use emissions	Kyoto gases
IMACLIM	CIRED, France	General	Recursive dynamic	2100	Medium	None	CO_2 from fossil fuel combustion and industry
IMAGE	PBL, the Netherlands	Partial	Recursive dynamic	2100	High	Endogenous land use dynamics	All GHGs and other radiative agents
MERGE-ETL	PSI, Switzerland	General	Intertemporal optimization	2100	High	MACs for land use emissions	All GHGs and other radiative agents
MESSAGE-MACRO	IIASA, Austria	General	Intertemporal optimization	2100	High	MACs for land use emissions, afforestation	All GHGs and other radiative agents
POLES	JRC IPTS, EU / EDDEN, France	Partial	Recursive dynamic	2100	High	None	Kyoto gases from fossil fuel combustion and industry
REMIND	PIK, Germany	General	Intertemporal optimization	2100	High	MACs for land use emissions	All GHGs and other radiative agents
WITCH	FEEM, Italy	General	Intertemporal optimization	2100	Low	MACs for land use emissions	Kyoto gases

* Many of the global IAMs participating in this study can also be run coupled with land use models to endogenize land use dynamics.

TABLE S3. GLOBAL IAMS PARTICIPATING IN WILKERSON, LEIBOWICZ, TURNER AND WEYANT (102) MODEL DIAGNOSTIC STUDY. SOURCE: (102). NOTE: MAC = MARGINAL ABATEMENT COSTS.

Global IAM	Modelling team home institute	Equil-ibrium type	Modelling approach	Time hor-izon	Resolution of energy supply	Representation of land use *	Coverage of greenhouse gases
EPPA	MIT, USA	General	Recursive dynamic	2100	Medium	MACs for land use emissions	All GHGs and other radiative agents
GCAM	PNNL, USA	Partial	Recursive dynamic	2100	High	Endogenous land use dynamics, afforestation	All GHGs and other radiative agents
MERGE	EPRI, USA	General	Intertemporal optimization	2100	High	MACs for land use emissions	All GHGs and other radiative agents

* Many of the global IAMs participating in this study can also be run coupled with land use models to endogenize land use dynamics.

Detailed model-specific documentation for each model in Tables S2 and S3 is referenced below. In many cases, original models have been subsequently developed into variants (e.g., with or without certain elements) and run under study-specific parameterizations.

Standardized model reference cards and documentation with easily comparable information are also available online for many of these models: https://wiki.ucl.ac.uk/display/ADVIAM/Review

AIM-Enduse: Hibino, G., R. Pandey, Y. Matsuoka, M. Kainuma (2013). A Guide to AIM-Enduse Model. National Institute of Environmental Studies, Japan. http://www.nies.go.jp/gaiyo/media_kit/16.AIM/Enduse/manual.html,

DNE21+: Sano, F., K. Akimoto, T. Homma, J. Oda, K. Wada (2012). Analysis of Asian long-term climate change mitigation in power generation sector. 3rd IAEE Asian Conference, Kyoto, Japan. http://eneken.ieej.or.jp/3rd_IAEE_Asia/pdf/paper/044p.pdf

EPPA: Paltsev, S., J. Reilly, H. Jacoby, R. Eckaus, J. McFarland, M. Sarofim, M. Asadoorian and M. Babiker (2005). The MIT Emissions Prediction and Policy Analysis (EPPA) Model: Version 4. Boston, MA, MIT Joint Program on the Science and Policy of Global Change.

GCAM: www.globalchange.umd.edu/models/gcam

GEM-E3: www.gem-e3.net

IMACLIM: Waisman, H., C. Guivarch, F. Grazi, J-C. Hourcade (2012). The Imaclim-R Model: Infrastructures, Technical Inertia and the Costs of Low Carbon Futures under Imperfect Foresight. Climatic Change 114(1): 101-20.

IMAGE: Stehfest, E., D. van Vuuren, T. Kram, L. Bouwman, R. Alkemade, M. Bakkenes, H. Biemans, A. Bouwman, M. den Elzen, J. Janse, P. Lucas, J. van Minnen, M. Müller and A. Prins (2014). Integrated Assessment of Global Environmental Change with IMAGE 3.0. Model description and policy applications. The Hague, Netherlands Environmental Assessment Agency (PBL).

MERGE: Manne, A. & R. Richels. (2004). MERGE: An Integrated Assessment Model for Global Climate Change. Stanford University, Stanford, CA. www.stanford.edu/group/MERGE/GERAD1.pdf

MERGE-ETL: Marcucci, A. & H. Turton (2012) The MERGE-ETL Model Documentation. Paul Scherrer Institute, Villingen. www.psi.ch/eem/ModelsEN/2012MergeDescription.pdf

MESSAGE: https://wiki.ucl.ac.uk/display/ADVIAM/MESSAGE

POLES: Institute for Prospective Technological Studies (2010). Prospective Outlook on Long-Term Energy Systems, POLES Manual, Version 6.1. European Commission Joint Research Centre. http://ipts.jrc.ec.europa.eu/activities/energy-and-transport/documents/POLESdescription.pdf

REMIND: Luderer, G., M. Leimbach, N. Bauer, E. Kriegler, L. Baumstark, C. Bertram, A. Giannousakis, J. Hilaire, D. Klein, A. Levesque, I. Mouratiadou, M. Pehl, R. Pietzcker, F. Piontek, N. Roming, A. Schultes, V.J. Schwanitz, J. Strefler (2013). Description of the REMIND Model (Version 1.6). Social Science Research Network (SSRN), Rochester, NY . Available at SSRN: http://ssrn.com/abstract=2697070 and www.pik-potsdam.de/research/sustainable- solutions/models/remind

WITCH: Bosetti, V., C. Carraro, M. Galeotti, E. Massetti, M. Tavoni (2006). WITCH: A World Induced Technical Change Hybrid Model. The Energy Journal 27: 13-38.

(5) Examples of global sensitivity analyses by process-based IAMs.

Local methods (also 'OAT' or one-at-a-time) test output sensitivities to changes in single inputs; global methods vary all other inputs as well. Global methods include screening methods (e.g., Morris) and variance-based methods (e.g., Sobol). This section provides illustrative examples of global methods applied to IAMs.

van der Sluijs, *et al.* (147) use the Morris method for global sensitivity analysis to explore quantitative uncertainties in parameter values in the TIMER energy model which is part of the IMAGE integrated assessment framework. 300 variables were varied over a range 0.5 to 1.5 times the default values. Input variables and model components to which projected CO_2 emissions were most sensitive included: economic activity; rates of technological improvement; supply curves for fossil fuels and renewables; rates of energy efficiency improvement. Campolongo and Braddock (148) use the full IMAGE model to test the sensitivity of atmospheric CO_2 concentrations, temperature change in the mixed ocean layer, and sea level rise to both the main effects and interaction effects of six input parameters including biotic growth, climate sensitivity, and rate of increase in net primary production.

Branger, Giraudet, Guivarch and Quirion (149) test the sensitivity of the Res-IRF model of energy demand in the residential sector in France. Using the Morris method applied to the full set of inputs and parameters, they found that sectoral energy demand was most sensitive to exogenously-specified future energy prices and the parameterization of energy service elasticity. Bosetti, *et al.* (92) test the sensitivity of two global energy-economy models, GCAM and WITCH, to uncertain technology cost and performance assumptions. Using different global methods applied to five energy technologies, they found emission outcomes were strongly sensitive to the costs of nuclear power, but that varying model responses to technology cost sensitivities were attributable to structural differences between models. McJeon, *et al.* (108) similarly generate a large number of scenarios using the GCAM model by varying combinations of technology cost assumptions. By apply a scenario discovery method to identify combinations that result in mitigation costs exceeding a threshold (defined as the 80^{th} percentile cost), they found that future costs of carbon capture and storage (CCS) were most influential over model outcomes. Pye, Sabio and Strachan (111) run a global sensitivity analysis on the ESME model which uses probability distributions of key inputs. They find the cost of a future UK energy transition consistent with climate change targets is most sensitive to assumptions on gas prices and biomass availability.

References

1. 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*, eds Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, & Midgley PM (Cambridge University Press, Cambridge, UK and New York, USA.).
2. Collins M, *et al.* (2013) Long-term Climate Change: Projections, Commitments and Irreversibility. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, & Midgley PM (Cambridge University Press, Cambridge, UK and New York, USA.).
3. Clarke L, *et al.* (2014) Chapter 6: Assessing Transformation Pathways. *Working Group III contribution to the IPCC 5th Assessment Report, Climate Change 2014: Mitigation of Climate Change*, (Cambridge University Press, Cambridge, UK and New York, NY, USA).
4. Sathaye J & Shukla PR (2013) Methods and Models for Costing Carbon Mitigation. *Annual Review of Environment and Resources* 38(1):137-168.
5. Krey V, *et al.* (2014) Annex II: Metrics & Methodology. *Working Group III contribution to the IPCC 5th Assessment Report, Climate Change 2014: Mitigation of Climate Change*, (Cambridge University Press, Cambridge, UK and New York, NY, USA).
6. Moore FC & Diaz DB (2015) Temperature impacts on economic growth warrant stringent mitigation policy. *Nature Clim. Change* 5(2):127-131.
7. Stern N (2006) *The Stern Review on the Economics of Climate Change* (Cambridge University Press, Cambridge, UK).
8. Millner A & McDermott TKJ (2016) Model confirmation in climate economics. *Proceedings of the National Academy of Sciences* 113(31):8675-8680.
9. UK_CCC (2013) Fourth Carbon Budget Review. (UK Committee on Climate Change, London, UK).
10. Edenhofer O & Minx J (2014) Mapmakers and navigators, facts and values. *Science* 345(6192):37-38.
11. IEA (2014) *World Energy Outlook Special Report: Redrawing the Energy-Climate Map* (International Energy Agency, Paris, France).
12. UNFCCC (2015) Synthesis report on the aggregate effect of the intended nationally determined contributions. (United Nations Framework Convention on Climate Change (UNFCCC) Secretariat, Bonn, Germany).
13. UNEP (2015) The Emissions Gap Report 2015. (United Nations Environment Programme (UNEP), Nairobi, Kenya).
14. Riahi K, *et al.* (2012) Energy Pathways for Sustainable Development. *The Global Energy Assessment.*, (Cambridge University Press, Cambridge, UK).
15. Rogelj J, McCollum DL, & Riahi K (2013) The UN's 'Sustainable Energy for All' initiative is compatible with a warming limit of 2oC. *Nature Clim. Change* 3(6):545-551.
16. Stechow Cv, *et al.* (2015) Integrating Global Climate Change Mitigation Goals with Other Sustainability Objectives: A Synthesis. *Annual Review of Environment and Resources* 40(1):363-394.
17. van Vuuren DP, *et al.* (2015) Pathways to achieve a set of ambitious global sustainability objectives by 2050: Explorations using the IMAGE integrated assessment model. *Technological Forecasting and Social Change* 98:303-323.
18. Barlas Y & Carpenter S (1990) Philosophical roots of model validation: Two paradigms. *System Dynamics Review* 6(2):148-166.
19. Peace J & Weyant J (2008) Insights Not Numbers: The Appropriate Use of Economic Models. (Pew Center on Global Climate Change, Washington, DC).
20. Kunreuther H, *et al.* (2014) Integrated Risk and Uncertainty Assessment of Climate Change Response Policies. *Working Group III contribution to the IPCC 5th Assessment Report, Climate Change 2014: Mitigation of Climate Change*, eds Edenhofer O, Pichs-Madruga R, Sokona Y, Farahani E, Kadner S, Seyboth K, Adler A, Baum I, Brunner S, Eickemeier P, *et al.* (Cambridge University Press, Cambridge, UK and New York, NY, USA).
21. Liu J, *et al.* (2015) Systems integration for global sustainability. *Science* 347(6225).

22. Iyer GC, Clarke LE, Edmonds JA, Hultman NE, & McJeon HC (2015) Long-term payoffs of near-term low-carbon deployment policies. *Energy Policy* 86:493-505.
23. Vuuren DPv, *et al.* (2016) Carbon budgets and energy transition pathways. *Environmental Research Letters* 11(7):075002.
24. van der Sluijs JP, Petersen AC, Janssen PHM, Risbey JS, & Ravetz JR (2008) Exploring the quality of evidence for complex and contested policy decisions. *Environmental Research Letters* 3(2):024008.
25. Beck M & Krueger T (2016) The epistemic, ethical, and political dimensions of uncertainty in integrated assessment modeling. *Wiley Interdisciplinary Reviews: Climate Change*:n/a-n/a.
26. Rosen RA & Guenther E (2015) The economics of mitigating climate change: What can we know? *Technological Forecasting and Social Change* 91(0):93-106.
27. Smith SJ, *et al.* (2015) Long history of IAM comparisons. *Nature Climate Change* 5(5):391.
28. Flato GM (2011) Earth system models: an overview. *Wiley Interdisciplinary Reviews: Climate Change* 2(6):783-800.
29. Moss RH, *et al.* (2010) The next generation of scenarios for climate change research and assessment. *Nature* 463:747-756.
30. NAS (2016) Assessment of Approaches to Updating the Social Cost of Carbon: Phase 1 Report on a Near-Term Update. (Committee on Assessing Approaches to Updating the Social Cost of Carbon, Board on Environmental Change and Society, National Academies of Sciences, Engineering, and Medicine, Washington, DC), pp 1–74.
31. Greenstone M, Kopits E, & Wolverton A (2013) Developing a Social Cost of Carbon for US Regulatory Analysis: A Methodology and Interpretation. *Review of Environmental Economics and Policy* 7(1):23-46.
32. Oreskes N (1998) Evaluation (not validation) of quantitative models. *Environ. Health Perspect.* 106(6):1453–1460.
33. Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. *System Dynamics Review* 12(3):183-210.
34. Oreskes N, Shrader-Frechette K, & Belitz K (1994) Verification, validation, and confirmation of numerical models in the earth sciences. *Science* 263(5147):641–646.
35. Sargent RG (2013) Verification and validation of simulation models. *J of Sim* 7(1):12-24.
36. Schwanitz VJ (2013) Evaluating integrated assessment models of global climate change. *Environmental Modelling & Software* 50(0):120-131.
37. Pirtle Z, Meyer R, & Hamilton A (2010) What does it mean when climate models agree? A case for assessing independence among general circulation models. *Environmental Science & Policy* 13:351–361.
38. van der Sluijs JP, *et al.* (2005) Combining Quantitative and Qualitative Measures of Uncertainty in Model-Based Environmental Assessment: The NUSAP System. *Risk Anal.* 25(2):481-492.
39. Oppenheimer M, Little CM, & Cooke RM (2016) Expert judgement and uncertainty quantification for climate change. *Nature Clim. Change* 6(5):445-451.
40. Schneider SH (1997) Integrated assessment modeling of global climate change: transparent rational tool for policy making or opaque screen hiding valueladen assumptions? *Environmental Modelling and Assessment* 2:229-249.
41. Funtowicz SO & Ravetz JR (1993) Science for the post-normal age. *Futures* 25(7):735–755.
42. Tebaldi C & Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 365(1857):2053-2075.
43. Stainforth DA, Allen MR, Tredger ER, & Smith LA (2007) Confidence, uncertainty and decision-support relevance in climate predictions. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 365(1857):2145-2161.
44. van Ruijven B, *et al.* (2010) Uncertainty from Model Calibration: Applying a New Method to Transport Energy Demand Modelling. *Environ Model Assess* 15(3):175-188.
45. Oreskes N & Belitz K (2001) Philosophical Issues in Model Assessment. *Model Validation: Perspectives in Hydrological Science*, eds Anderson MG & Bates PD (John Wiley & Sons, Chichester, UK), pp 23-41.
46. Beven K (2006) A manifesto for the equifinality thesis. *Journal of Hydrology* 320(1–2):18–36.
47. Beugin D & Jaccard M (2012) Statistical Simulation to Estimate Uncertain Behavioral Parameters of Hybrid Energy-Economy Models. *Environ Model Assess* 17(1-2):77-90.

48. Mauritsen T, *et al.* (2012) Tuning the climate of a global model. *Journal of Advances in Modeling Earth Systems* 4(3).
49. Schindler DE & Hilborn R (2015) Prediction, precaution, and policy under global change. *Science* 347(6225):953-954.
50. Parker WS (2013) Ensemble modeling, uncertainty and robust predictions. *Wiley Interdisciplinary Reviews: Climate Change* 4(3):213-223.
51. DeCarolis JF, Hunter K, & Sreepathi S (2012) The case for repeatable analysis with energy economy optimization models. *Energy Econ.* 34:1845-1853.
52. Weyant J (2009) A perspective on integrated assessment: an editorial comment. *Climatic Change* 95:317-323.
53. Trutnevyte E (2016) Does cost optimisation approximate the real-world energy transition? *Energy* 106:182-193.
54. Rosen RA & Guenther E (2016) The energy policy relevance of the 2014 IPCC Working Group III report on the macro-economics of mitigating climate change. *Energy Policy* 93:330-334.
55. Stocker TF, D. Qin, G.-K. Plattner, L.V. Alexander, S.K. Allen, N.L. Bindoff, F.-M. Bréon, J.A. Church, U. Cubasch, S. Emori, P. Forster, P. Friedlingstein, N. Gillett, J.M. Gregory, D.L. Hartmann, E. Jansen, B. Kirtman, R. Knutti, K. Krishna Kumar, P. Lemke, J. Marotzke, V. Masson-Delmotte, G.A. Meehl, I.I. Mokhov, S. Piao, V. Ramaswamy, D. Randall, M. Rhein, M. Rojas, C. Sabine, D. Shindell, L.D. Talley, D.G. Vaughan, S.-P. Xie (2013) Technical Summary. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, & Midgley PM (Cambridge University Press, Cambridge, UK and New York, USA.).
56. Chaturvedi V, *et al.* (2013) Model evaluation and hindcasting: An experiment with an integrated assessment model. *Energy* 61:479-490.
57. Guivarch C, Hallegatte S, & Crassous R (2009) The resilience of the Indian economy to rising oil prices as a validation test for a global energy–environment–economy CGE model. *Energy Policy* 37(11):4259-4266.
58. DeCarolis JF (2011) Using modeling to generate alternatives (MGA) to expand our thinking on energy futures. *Energy Econ.* 33(2):145-152.
59. van Vuuren DP, *et al.* (2010) What do near-term observations tell us about long-term developments in greenhouse gas emissions? *Global Environmental Change* 103:635-642.
60. van Ruijven B, *et al.* (2011) Model projections for household energy use in India. *Energy Policy* 39:7747-7761.
61. Macknick J (2011) Energy and CO2 emission data uncertainties. *Carbon Management* 2(2):189-205.
62. Van Minnen J, *et al.* (2009) The importance of three centuries of land-use change for the global and regional terrestrial carbon cycle. *Climatic Change* 97(1-2):123-144.
63. Cubasch U, *et al.* (2013) Introduction. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, & Midgley PM (Cambridge University Press, Cambridge, UK and New York, USA.).
64. Nakicenovic N, *et al.* (2000) *Special Report on Emissions Scenarios* (Cambridge University Press, Cambridge, UK).
65. van Vuuren DP & O'Neill BC (2006) The Consistency of IPCC's SRES Scenarios to 1990–2000 Trends and Recent Projections. *Climatic Change* 75(1-2):9-46.
66. Manning MR, *et al.* (2010) Misrepresentation of the IPCC CO2 emission scenarios. *Nature Geoscience* 3(6):376-377.
67. Raupach ME, *et al.* (2007) Global and regional drivers of accelerating CO2 emissions. *Proceedings of the National Academy of Sciences* 104(24):10288-10293.
68. Peters GP, *et al.* (2012) The challenge to keep global warming below 2oC. *Nature Clim. Change* 3:4-6.
69. Schwanitz VJ & Wierling A (2016) Offshore wind investments – Realism about cost developments is necessary. *Energy* 106:170-181.
70. Craig PP, Gadgil A, & Koomey JG (2002) What Can History Teach Us? A Retrospective Examination of Long-Term Energy Forecasts for the United States. *Annual Review of Energy and the Environment* 27(1):83-118.

71. Smil V (2000) Perils of long-range energy forecasting: reflections on looking far ahead. *Technological Forecasting and Social Change* 65(3):251–264.
72. Koomey J, Craig P, Gadgil A, & Lorenzetti D (2003) Improving Long-Range Energy Modeling: A Plea for Historical Retrospectives. *The Energy Journal* 24(4):75-92.
73. Schmidt GA, Shindell DT, & Tsigaridis K (2014) Reconciling warming trends. *Nature Geosci* 7(3):158-160.
74. Kaldor N (1957) A Model of Economic Growth. *The Economic Journal* 67(268):591–624.
75. Leimbach M, Kriegler E, Roming N, & Schwanitz J (2015) Future growth patterns of world regions – A GDP scenario approach. *Global Environmental Change*.
76. Hargreaves JC & Annan JD (2014) Can we trust climate models? *Wiley Interdisciplinary Reviews: Climate Change* 5(4):435-440.
77. van Ruijven B, *et al.* (2008) Modeling Energy and Development: An Evaluation of Models and Concepts. *World Devel.* 36(12):2801-2821.
78. Wilson C, Grubler A, Bauer N, Krey V, & Riahi K (2012) Future capacity growth of energy technologies: are scenarios consistent with historical evidence? *Climatic Change* 118(2):381-395.
79. Iyer G, *et al.* (2015) Diffusion of low-carbon technologies and the feasibility of long-term climate targets. *Technological Forecasting and Social Change* 90:103-118.
80. van Sluisveld M, *et al.* (2015) Comparing future patterns of energy system change in 2°C scenarios with historically observed rates of change. *Global Environmental Change* 35:436-449.
81. Jakeman AJ, Letcher RA, & Norton JP (2006) Ten iterative steps in development and evaluation of environmental models. *Environmental Modelling & Software* 21(5):602-614.
82. Crout NMJ, Tarsitano D, & Wood AT (2009) Is my model too complex? Evaluating model formulation using model reduction. *Environmental Modelling & Software* 24(1):1-7.
83. Stocker T (2011) Model Hierarchy and Simplified Climate Models. *Introduction to Climate Modelling*, Advances in Geophysical and Environmental Mechanics and Mathematics, (Springer Berlin Heidelberg), pp 25-51.
84. Held IM (2005) The Gap between Simulation and Understanding in Climate Modeling. *Bulletin of the American Meteorological Society* 86(11):1609-1614.
85. van Vuuren DP, *et al.* (2011) How well do integrated assessment models simulate climate change? *Climatic Change* 104(2):255-285.
86. Baldos ULC & Hertel TW (2013) Looking back to move forward on model validation: insights from a global model of agricultural land use. *Environmental Research Letters* 8(3):034024.
87. Treut HL, *et al.* (2007) Historical Overview of Climate Change. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, eds Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, & Miller HL (Cambridge University Press, Cambridge, UK and New York, NY, USA), Vol I.
88. Taylor KE, Stouffer RJ, & Meehl GA (2011) An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society* 93(4):485-498.
89. Gaskins DW, Jr. & Weyant JP (1993) Model Comparisons of the Costs of Reducing CO2 Emissions. *The American Economic Review* 83(2):318-323.
90. Tavoni M, *et al.* (2015) Post-2020 climate agreements in the major economies assessed in the light of global models. *Nature Clim. Change* 5(2):119-126.
91. Clarke L, *et al.* (2009) International climate policy architectures: Overview of the EMF 22 International Scenarios. *Energy Econ.* 31(Supplement 2: (International, U.S. and E.U. Climate Change Control Scenarios: Results from EMF 22)):S64-S81.
92. Bosetti V, *et al.* (2015) Sensitivity to energy technology costs: A multi-model comparison analysis. *Energy Policy* 80(0):244-263.
93. 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. *Technological Forecasting and Social Change* 90:8-23.
94. Kriegler E, *et al.* (2016) Will economic growth and fossil fuel scarcity help or hinder climate stabilization? *Climatic Change* 136(1):7-22.
95. Kriegler E, *et al.* (2014) What does the 2C target imply for a global climate agreement in 2020? The LIMITS study on Durban Platform scenarios. *Climate Change Economics*.
96. Yokohata T, *et al.* (2010) Structural similarities and differences in climate responses to CO2 increase between two perturbed physics ensembles. *Journal of Climate* 23(6):1392-1410.

97. Morgan MG & Keith D (2008) Improving the way we think about projecting future energy use and emissions of carbon dioxide. *Climatic Change* 90:189-215.
98. Tavoni M & Tol RSJ (2010) Counting only the hits? The risk of underestimating the costs of stringent climate policy. *Climatic Change* 100(3):769-778.
99. Bennett ND, *et al.* (2013) Characterising performance of environmental models. *Environmental Modelling & Software* 40:1-20.
100. Andrews T, Gregory JM, Webb MJ, & Taylor KE (2012) Forcing, feedbacks and climate sensitivity in CMIP5 coupled atmosphere-ocean climate models. *Geophysical Research Letters* 39(9):n/a-n/a.
101. Kriegler E, *et al.* (2015) Diagnostic indicators for integrated assessment models of climate policy. *Technological Forecasting and Social Change* 90:45-61.
102. Wilkerson JT, Leibowicz BD, Turner DD, & Weyant JP (2015) Comparison of integrated assessment models: Carbon price impacts on U.S. energy. *Energy Policy* 76:18-31.
103. Saltelli A, *et al.* (2008) *Global Sensitivity Analysis: The Primer.* (Wiley, Chichester, England).
104. Stainforth DA, *et al.* (2005) Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* 433:403–406.
105. Rowlands DJ, *et al.* (2012) Broad range of 2050 warming from an observationally constrained large climate model ensemble. *Nature Geosciences* 5(4):256-260.
106. van Vuuren DP, de Vries B, Beusen A, & Heuberger PSC (2008) Conditional probabilistic estimates of 21st century greenhouse gas emissions based on the storylines of the IPCC-SRES scenarios. *Global Environmental Change* 18(4):635-654.
107. Koelbl BS, van den Broek MA, van Ruijven BJ, Faaij APC, & van Vuuren DP (2014) Uncertainty in the deployment of Carbon Capture and Storage (CCS): A sensitivity analysis to techno-economic parameter uncertainty. *International Journal of Greenhouse Gas Control* 27:81-102.
108. McJeon HC, *et al.* (2011) Technology interactions among low-carbon energy technologies: What can we learn from a large number of scenarios? *Energy Econ.* 33:619-631.
109. Saltelli A & D'Hombres B (2010) Sensitivity analysis didn't help. A practitioner's critique of the Stern review. *Global Environmental Change* 20(2):298-302.
110. Borgonovo E (2010) Sensitivity analysis with finite changes: An application to modified EOQ models. *European Journal of Operational Research* 200(1):127-138.
111. Pye S, Sabio N, & Strachan N (2015) An integrated systematic analysis of uncertainties in UK energy transition pathways. *Energy Policy* 87:673-684.
112. Mundaca L, Neij L, Worrell E, & McNeil M (2010) Evaluating Energy Efficiency Policies with Energy-Economy Models. *Annual Review of Environment and Resources* 35(1):305-344.
113. NCC (2015) IAM helpful or not? *Nature Clim. Change* 5(2):81-81.
114. Voosen P (2016) Climate scientists open up their black boxes to scrutiny. *Science* 354(6311):401.
115. Messner S & Strubegger M (1995) User's guide for MESSAGE III. (IIASA, Laxenburg, Austria).
116. Alcamo J, Kreileman GJJ, Krol MS, & Zuidema G (1994) Modeling the global society-biosphere-climate system: Part 1: Model description and testing. *Water. Air. Soil Pollut.* 76(1):1-35.
117. Gargiulo M & Gallachóir BÓ (2013) Long-term energy models: Principles, characteristics, focus, and limitations. *Wiley Interdisciplinary Reviews: Energy and Environment* 2(2):158-177.
118. McDowall W, Trutnevyte E, Tomei J, & Keppo I (2014) Reflecting on Scenarios. (UKERC (UK Energy Research Centre) Energy Systems Theme, London, UK).
119. Harmsen MJHM, *et al.* (2015) How well do integrated assessment models represent non-CO2 radiative forcing? *Climatic Change* 133:565-582.
120. Nakicenovic N, Lempert RJ, & Janetos AC (2014) A Framework for the Development of New Socio-economic Scenarios for Climate Change Research: Introductory Essay. *Climatic Change* 122(3):351-361.
121. Fischhoff B (2015) The realities of risk-cost-benefit analysis. *Science* 350(6260).
122. Kloprogge P, Sluijs JP, & Wardekker JA (2007) Uncertainty communication: issues and good practice. (Copernicus Institute for Sustainable Development and Innovation, Utrecht University Utrecht, the Netherlands).
123. Cooke RM (2015) Messaging climate change uncertainty. *Nature Clim. Change* 5(1):8-10.

124. Lempert RJ (2015) Climate targets: Values and uncertainty. *Nature Clim. Change* 5(10):914-915.
125. Drouet L, Bosetti V, & Tavoni M (2015) Selection of climate policies under the uncertainties in the Fifth Assessment Report of the IPCC. *Nature Clim. Change* 5(10):937-940.
126. Krueger T, Page T, Hubacek K, Smith L, & Hiscock K (2012) The role of expert opinion in environmental modelling. *Environmental Modelling & Software* 36:4-18.
127. Burke M, *et al.* (2016) Opportunities for advances in climate change economics. *Science* 352(6283):292.
128. Nordhaus WD (2013) *The climate casino: Risk, uncertainty, and economics for a warming world* (Yale University Press, New Haven, CT).
129. Hope C & Hope M (2013) The social cost of CO2 in a low-growth world. *Nature Clim. Change* 3(8):722-724.
130. Luderer G, *et al.* (2013) Economic mitigation challenges: how further delay closes the door for achieving climate targets. *Environmental Research Letters* 8(3):034033.
131. Stanton EA, Ackerman F, & Kartha S (2009) Inside the integrated assessment models: four issues in climate economics. *Climate and Development* 1:166-184.
132. Ackerman F, DeCanio SJ, Howarth RB, & Sheeran K (2009) Limitations of integrated assessment models of climate change. *Climatic Change* 95(3):297-315.
133. Pindyck RS (2013) Climate Change Policy: What Do the Models Tell Us? *J. Econ. Lit.* 51(3):860-872.
134. Weitzman ML (2009) On modelling and interpreting the economics of catastrophic climate change. *The Review of Economics and Statistics* 91(1):1–19.
135. Cai Y, Judd KL, Lenton TM, Lontzek TS, & Narita D (2015) Environmental tipping points significantly affect the cost−benefit assessment of climate policies. *Proceedings of the National Academy of Sciences* 112(15):4606-4611.
136. Lontzek TS, Cai Y, Judd KL, & Lenton TM (2015) Stochastic integrated assessment of climate tipping points indicates the need for strict climate policy. *Nature Clim. Change* 5(5):441-444.
137. Stern N (2016) Current climate models are grossly misleading. *Nature* 530:407-409.
138. Metcalf G & Stock J (2015) The Role of Integrated Assessment Models in Climate Policy: A User's Guide and Assessment. (The Harvard Project on Climate Agreements, Cambrige, MA).
139. Arent DJ, *et al.* (2014) Key economic sectors and services. *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel of Climate Change*, eds Field CB, Barros VR, Dokken DJ, Mach KJ, Mastrandrea MD, Bilir TE, Chatterjee M, Ebi KL, Estrada YO, Genova RC, *et al.* (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA), pp 659-708.
140. Beckman J, Hertel T, & Tyner W (2011) Validating energy-oriented CGE models. *Energy Econ.* 33(5):799-806.
141. van Ruijven B, de Vries B, van Vuuren DP, & van der Sluijs JP (2010) A global model for residential energy use: Uncertainty in calibration to regional data. *Energy* 35(1):269-282.
142. Fujimori S, Dai H, Masui T, & Matsuoka Y (2016) Global energy model hindcasting. *Energy* 114:293-301.
143. Jones CI & Romer PM (2010) The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital. *American Economic Journal: Macroeconomics* 2(1):224-245.
144. Grubler A, *et al.* (2012) Energy Primer. *Global Energy Assessment*, (Cambridge University Press, Cambridge, UK).
145. Wilson C (2016) Long-term dynamics of technological change in mitigation scenarios. in *International Energy Workshop* (Cork, Ireland).
146. Kriegler E, *et al.* (2015) Making or breaking climate targets: The AMPERE study on staged accession scenarios for climate policy. *Technological Forecasting and Social Change* 90, Part A(0):24-44.
147. van der Sluijs JP, *et al.* (2002) Uncertainty Assessment of the IMAGE/TIMER B1 CO2 Emissions Scenario Using the NUSAP Method. (Dutch National Research Program on Climate Change, Bilthoven, the Netherlands).
148. Campolongo F & Braddock R (1999) Sensitivity analysis of the IMAGE Greenhouse model. *Environmental Modelling & Software* 14(4):275-282.

149. Branger F, Giraudet L-G, Guivarch C, & Quirion P (2015) Global sensitivity analysis of an energy–economy model of the residential building sector. *Environmental Modelling & Software* 70:45-54.