Climatic Change (2021) 166:3
https://doi.org/10.1007/s10584-021-03099-9

Evaluating process-based integrated assessment models of climate change mitigation

Charlie Wilson[1,2] · Céline Guivarch[3] · Elmar Kriegler[4,5] · Bas van Ruijven[2] · Detlef P. van Vuuren[6,7] · Volker Krey[2] · Valeria Jana Schwanitz[8,9] · Erica L. Thompson[10]

Received: 12 August 2019 / Accepted: 19 April 2021/Published online: 04 May 2021
© The Author(s) 2021

Abstract
Process-based integrated assessment models (IAMs) project long-term transformation pathways in energy and land-use systems under what-if assumptions. IAM evaluation is necessary to improve the models' usefulness as scientific tools applicable in the complex and contested domain of climate change mitigation. We contribute the first comprehensive synthesis of process-based IAM evaluation research, drawing on a wide range of examples across six different evaluation methods including historical simulations, stylised facts, and model diagnostics. For each evaluation method, we identify progress and milestones to date, and draw out lessons learnt as well as challenges remaining. We find that each evaluation method has distinctive strengths, as well as constraints on its application. We use these insights to propose a systematic evaluation framework combining multiple methods to establish the appropriateness, interpretability, credibility, and relevance of process-based IAMs as useful scientific tools for informing climate policy. We also set out a programme of evaluation research to be mainstreamed both within and outside the IAM community.

Keywords Process-based integrated assessment model · IAM · Evaluation · Climate mitigation

1 Introduction

Process-based integrated assessment models (IAMs) represent linkages and trade-offs between energy, land use, climate, economy, and development (Sathaye and Shukla 2013; von Stechow et al. 2015). IAMs are useful for analysing long-term global climate change mitigation in a sustainable development context under what-if assumptions about future drivers of change. The IPCC's fifth assessment report distilled insights from 1134 scenarios from 30 global IAMs (Clarke et al. 2014; Krey et al. 2014). The more recent 2018

✉ Charlie Wilson
charlie.wilson@uea.ac.uk

Extended author information available on the last page of the article

IPCC special report on global warming of 1.5 °C was informed by 411 scenarios from 10 global IAMs (Rogelj et al. 2018). Process-based IAMs are also used more directly in climate policy formulation, including the periodic global stocktake of progress under the Paris Agreement (Grassi et al. 2018), international negotiations under the UNFCCC (UNEP 2015; UNFCCC 2015), and national strategies, targets, and regulatory appraisals (BEIS 2018; Weitzel et al. 2019).

In order to be useful scientific tools for climate policy analysis, policymakers need to have confidence in IAMs and their analyses. Evaluating modelling tools such as IAMs means assessing both the models and their performance so as to articulate the grounds on which they can be declared good enough for their intended use (Oreskes 1998). Evaluation is a necessarily broad and open-ended process.

IAM evaluation has a long history in multi-model comparison projects (Huntington et al. 1982), but other evaluation methods have been implemented on a more ad-hoc basis by individual modelling teams with little community-wide coordination or consolidation. While important for tacit learning and model development, such evaluation practices are less impactful on the wider climate research community.

Process-based IAMs have also been criticised for a range of perceived failings, including technological hubris (Anderson and Peters 2016), omitted drivers of sociotechnical change (Geels et al. 2017), and understating future uncertainties (Rosen and Guenther 2015). As IAMs have become more widely used for informing policy (van Beek et al. 2020), the lack of a coherent and systematic approach to evaluation has become more conspicuous by its absence.

In response, we contribute the first synthesis of IAM evaluation research, drawing on a wide range of examples across six different evaluation methods: historical simulations, near-term observations, stylised facts, model hierarchies from simple to complex, model inter-comparison projects (including diagnostic indicators), and sensitivity analysis. For each method, we review key milestones in historical development and application, and draw out lessons learnt as well as remaining challenges.

Following Cash et al. (2003), we also propose four criteria against which evaluation can help improve IAMs and their usefulness in policy contexts: appropriateness, interpretability, credibility, and relevance. We map each evaluation method onto these criteria, and conclude by arguing for a systematic evaluation framework which combines the strengths of multiple methods to overcome the limitations of any single method.

2 Process-based IAMs not benefit-cost models

Throughout this article, we use ‘IAMs’ to mean process-based integrated assessment models (or what Weyant (2017) calls ‘detailed process’ or DP IAMs). These IAMs:

(1) Represent explicitly the drivers and processes of change in global energy and land-use systems linked to the broader economy, often with a high degree of technological resolution in the energy supply
(2) Capture both biophysical and socioeconomic processes including human preferences, but do *not* generally include future impacts or damages of climate change on these processes
(3) Project cost-effective ‘optimal’ mitigation pathways under what-if assumptions or subject to pre-defined outcomes such as limiting global warming to 2 °C (Sathaye and Shukla 2013)

Many process-based IAMs originate in energy system models or energy-economy models which have since integrated land use, greenhouse gas emissions, and other climate-related processes (Krey et al. 2014; Sathaye and Shukla 2013). Examples of process-based IAMs include AIM-Enduse (Hibino et al. 2013), GCAM (Iyer et al. 2015), IMACLIM (Waisman et al. 2011), IMAGE (van Vuuren et al. 2015), MESSAGE-GLOBIOM (Krey et al. 2016), and REMIND (Luderer et al. 2013). We provide further examples in the Online Resources, including process flow diagrams of select IAMs in Online Resources 7.

In this article on model evaluation, we *do not* consider integrated assessment models used for cost-benefit analyses which have simplified representations of energy and land-use systems (and which are also referred to in the literature as 'IAMs'). Examples of cost-benefit integrated assessment models include DICE (Nordhaus 2013), PAGE (Hope and Hope 2013), and FUND (Anthoff and Tol 2013). These highly aggregated models are used in a cost-benefit framework to analyse economically optimal levels of abatement taking into account future impacts of climate change (Moore and Diaz 2015; Stern 2006). Cost-benefit models have also been widely applied in the USA and elsewhere for projecting the social cost of carbon in order to internalise climate-related impacts in regulatory appraisal processes (Greenstone et al. 2013; NAS 2016).

By focusing *only* on process-based IAMs, we follow established precedent which recognises fundamental differences between the two types of model (Kunreuther et al. 2014). Evaluation research confronts very different issues in process-based and cost-benefit models even though similar methods can be applied. For process-based IAMs with detailed representations of biophysical and socioeconomic processes across multiple subsystems, evaluation is concerned with how causal mechanisms and interactions generate outcomes of interest. As process-based IAMs are complex 'black boxes', running in specialised programming environments with high technical barriers to entry, evaluation is also concerned with issues of interpretability and transparency.

In contrast, causal processes encoded in cost-benefit type models are simplified and widely accessible (e.g. in spreadsheet tools) so understanding how outcomes of interest are generated is less of an issue. Rather, evaluation is concerned with the empirical and theoretical defensibility of assumptions such as discount rates (Metcalf and Stock 2015), of parameterised relationships such as climate impact damage functions (Cai et al. 2015), and of general modelling approaches such as the integration of mitigation co-benefits in a welfare maximisation framework (Stern 2016).

Consequently, we focus *only* on evaluation of process-based IAMs, and do *not* consider cost-benefit type models further.

3 What is IAM evaluation and why do it?

As Barlas and Carpenter (1990) observe: 'Models are not true or false but lie on a continuum of usefulness'. IAMs are useful for strengthening scientific understanding of coupled human and natural systems relevant to climate change (Moss et al. 2010). As an example, reference or baseline scenarios are modelled to characterise salient uncertainties in future development pathways (e.g. Riahi et al. 2017). IAMs are also useful for informing climate policymakers on the options and implications of decisions or indecision (Edenhofer and Minx 2014). As an example, policy or mitigation scenarios are modelled to help policymakers understand the consequences of different carbon pricing tariffs at regional or global scales (Vrontisi et al. 2018).

Evaluating IAMs should help improve their usefulness as scientific tools for policy-relevant analysis. Evaluation is an open-ended process testing both model structure and model behaviour (Barlas 1996; Eker et al. 2018).[1]

Evaluating model structure tests how the modelled system is represented in equations (encoding laws, principles, causal relationships, and drivers of change), parameterisations (making simplifying assumptions about complex phenomena), constraints (imposed as external conditions), variables, and values assigned to input variables or parameters (Pirtle et al. 2010).[2]

Evaluating model behaviour tests how observed system responses are reproduced. This is commonly done by selecting an evaluation period and tuning or calibrating specific model variables and parameters to match the initial conditions and external drivers of change observed over that period. The model is then run to test how well it predicts non-calibrated outcomes (Snyder et al. 2017).

IAMs represent complex, dynamic systems characterised by deep uncertainties. Uncertainties are *epistemic* (ignorance), *parametric* (inexactness), and *societal* (values) (van der Sluijs et al. 2008). 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 parameterisations (Oreskes 1998). Societal uncertainties are associated with values and worldviews which become embedded in model assumptions. Representation of social or economic processes not based on physical laws may be particularly contested (Oppenheimer et al. 2016; Schneider 1997).

As a result of these uncertainties, whether an IAM accurately represents the system being modelled cannot be definitively established (Oreskes et al. 1994; Sargent 2013). The same applies to other models used to assess environmental problems in coupled human-natural systems (Beck and Krueger 2016; van der Sluijs et al. 2005). As we discuss further below, this also means that for IAMs there is some overlap between uncertainty analysis and evaluation.

Evaluating IAM behaviour is similarly problematic. A close fit of IAM output to observational data does not necessarily mean the IAM accurately represents the modelled system. First, simulation results may be specific to the tuned parameterisations (Oreskes et al. 1994). Second, more than one model conceptualisation or parameterisation can generate the same output (Beugin and Jaccard 2012; van Ruijven et al. 2010). For example, van Ruijven et al. (2010) found multiple combinations of parameters could broadly reproduce historical transport energy demand in the TIMER model. Beugin and Jaccard (2012) similarly found multiple parameter distributions in backcasting simulations could reproduce observed technology adoption decisions in the CIMS model. Third, two or more settings (or errors) in the model

[1] The same emphasis on evaluating model structure and behaviour is found in climate modelling: 'Confidence in climate model projections is based on physical understanding of the climate system and its representation in climate models, and on a demonstration of how well models represent a wide range of processes and climate characteristics on various spatial and temporal scales' (p825, Flato et al. 2013).

[2] The term 'parameterisation' is used with two slightly different but related meanings. It can mean both a simplified representation of a complex process, or the assigning of a value to a parameter. As an example in IAMs, energy demand is responsive to energy prices for many reasons and competing causal mechanisms, but a model can represent this simply through an equation for demand as a function of price, with a value assigned to the price elasticity of demand (the percentage change in demand for a 1% change in price). The demand - price equation is a parameterisation of a complex process. And that equation is parameterised by assigning a value to the price elasticity coefficient.

inputs and parameterisations may partially cancel each other out (Schindler and Hilborn 2015).[3] All these evaluation issues apply generically to other models of complex systems.[4]

As a consequence of these difficulties in formally testing IAMs' structure and behaviour, IAM evaluation is necessarily a continual and iterative process of learning, development, and improvement (Barlas 1996).

We propose four inter-related criteria which guide this IAM evaluation process: appropriateness, interpretability, credibility, and relevance (see also Cash et al. (2003)).

First, IAM evaluation should improve the *appropriateness* of a model for addressing a specific scientific question (Jakeman et al. 2006; Sargent 2013). Matching model to task becomes harder as ever more extensive, higher resolution representations of coupled natural-human systems create IAMs with numerous possible applications (Gargiulo and Gallachóir 2013; van Beek et al. 2020).

Second, IAM evaluation should improve the *interpretability* of results, taking model structure and assumptions into account (DeCarolis et al. 2017; McDowall et al. 2014). Mapping meaning from the modelled world to the real world is of longstanding concern (Wynne 1984). A modelling mantra applicable to IAMs is to communicate 'insights not numbers' (Peace and Weyant 2008) alongside assumptions, uncertainties, and limitations (Beck and Krueger 2016; Cooke 2015; Kloprogge et al. 2007).

Third, IAM evaluation should improve the *credibility* of modelling analysis among user communities. How the producers and users of knowledge interact may be as important in determining the *credibility* of IAMs as the modelling analysis itself (Fischhoff 2015; Nakicenovic et al. 2014). In setting out best practice for model development and evaluation, Jakeman et al. (2006) urge a 'sceptical review' of models by users.

Fourth, IAM evaluation should improve the *relevance* of modelling analysis for informing scientific understanding and supporting decision-making on climate change mitigation. Cash et al. (2003) use the term 'salience' in a similar way to describe scientific information which is relevant to decision-making bodies or publics. The *credibility* and *relevance* criteria are most closely linked to the application of uncertain model results in complex and contested policy domains like climate change (Beck and Krueger 2016; van der Sluijs et al. 2008).

In strengthening the appropriateness, interpretability, credibility, and relevance of IAMs, it is important to emphasise 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 useful scientific tools for understanding mitigation pathways, decision options, and outcomes (Edenhofer and Minx 2014; Peace and Weyant 2008). IAMs sit alongside many other decision-support tools for informing climate policy, ranging from expert elicitations and bottom-up sectoral modelling to learning from experience and participatory appraisals (Kunreuther et al. 2014).

[3] The converse also holds. Divergence between model output and observational data does not necessarily mean the IAM is a poor representation of the modelled system. Divergence may be due to errors in inputs defining initial conditions or exogenous drivers of change, and large divergences may occur even when the model is only very slightly mis-specified (Thompson and Smith 2019).

[4] As an example in climate modelling, Tebaldi and Knutti (2007) note: 'Although model agreement with observations is very valuable in improving the model, and is a necessary condition for a model to be trusted, it does not definitely prove that the model is right for the right reason. There are well-known examples where errors in different components of a single model tend to cancel. The use of the same datasets for tuning and model evaluation raises the question of circular reasoning'.

4 Methods for evaluating process-based IAMs

IAMs began to emerge in the late 1980s, building on longer established traditions in energy system and macroeconomic modelling (as well as early system dynamics models). Concern for model evaluation dates back to these IAM precursors in the 1970s. In particular, the US-based Energy Modelling Forum (EMF) played a pivotal role in the early experimentation and development of methods now used in IAM evaluation (Smith et al. 2015).

These IAM evaluation methods can be classified into six types. Three evaluation methods use observational data: historical simulations, near-term observations, and stylised facts. Two evaluation methods use comparisons between models: model hierarchies from simple to complex, and model inter-comparison projects. Sensitivity analysis is a sixth evaluation method which is commonly applied to individual models, but can also form part of model inter-comparisons.

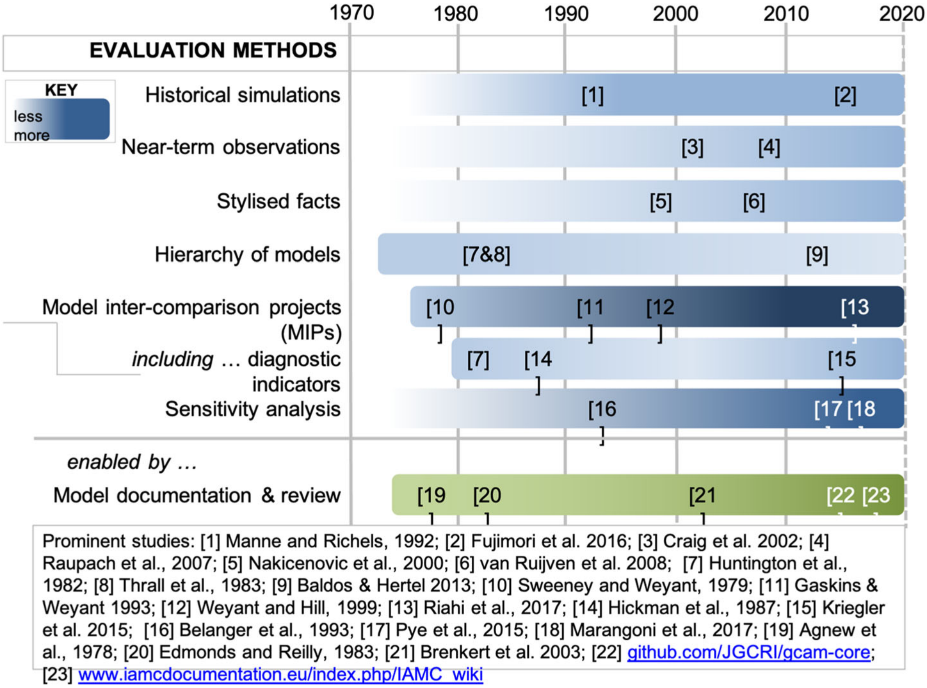
Sensitivity analysis is an example of how uncertainty analysis techniques overlap with evaluation methods given the many structural uncertainties which characterise IAMs' representation of incompletely understood socioeconomic and biophysical processes (Millner and McDermott 2016). Setting out the 'ten basic steps of good, disciplined model practice', Jakeman and colleagues emphasise the 'comprehensive testing of models' using methods to establish 'high enough confidence in estimates of model variables and parameters, taking into account the sensitivity of the outputs to all the parameters jointly, as well as the parameter uncertainties' (Jakeman et al. 2006) (see Online Resources 1 for further discussion).

For each of the six evaluation methods, we review historical development and applications (Fig. 1), and summarise lessons learnt as well as remaining challenges. We also consider the importance of model checks, documentation, and transparency for enabling independent verification of the scientific and policy applications of IAMs.

4.1 Historical simulations

Although prominent in early IAM evaluation practice (Toth 1994), historical simulations or hindcasting studies are relatively uncommon. Available simulations tend to be limited in time horizon, spatial scale, and model output compared to observations. Examples of simulated quantities in IAMs compared against observations include energy demand for the USA during 1960–1990 (Manne and Richels 1992); energy use in US buildings during the period 1995–2010 (Chaturvedi et al. 2013); the Indian economy's response to rising oil prices during 2003–2006 (Guivarch et al. 2009); and transportation energy demand in Western Europe during 1970–2003 (van Ruijven et al. 2010). In each case, the simulations led to revised modelling assumptions to reduce divergence from observations. (For further details and examples, see Online Resources 2.)

One hindcasting study with the AIM-CGE model compared a much broader set of simulated quantities to historical data including the primary energy and electricity supply mixes at both global and regional scales (Fujimori et al. 2016). An analogous study with the GCAM model examined model fit to historical global and regional land-use allocations (Snyder et al. 2017). These global hindcasting studies are rare as IAMs represent very diverse biophysical and socioeconomic processes as well as policy signals (van Vuuren et al. 2010). Simulated quantities must be sufficiently disaggregated to match this heterogeneity in underlying causal mechanisms (Schwanitz 2013). Relevant causal mechanisms (or model components) should also be structurally constant over the simulation period.



Prominent studies: [1] Manne and Richels, 1992; [2] Fujimori et al. 2016; [3] Craig et al. 2002; [4] Raupach et al., 2007; [5] Nakicenovic et al., 2000; [6] van Ruijven et al. 2008; [7] Huntington et al., 1982; [8] Thrall et al., 1983; [9] Baldos & Hertel 2013; [10] Sweeney and Weyant, 1979; [11] Gaskins & Weyant 1993; [12] Weyant and Hill, 1999; [13] Riahi et al., 2017; [14] Hickman et al., 1987; [15] Kriegler et al. 2015; [16] Belanger et al., 1993; [17] Pye et al., 2015; [18] Marangoni et al., 2017; [19] Agnew et al., 1978; [20] Edmonds and Reilly, 1983; [21] Brenkert et al. 2003; [22] github.com/JGCRI/gcam-core; [23] www.iamcdocumentation.eu/index.php/IAMC_wiki

Fig. 1 Historical development and landmark studies in IAM evaluation methods

The ability of IAMs to reproduce historical observations has further limitations as an evaluation method for several reasons.

First, historical simulations cannot demonstrate models' predictive reliability in future conditions that lie outside the range of historical experience (Oreskes 1998). This is a particular issue for IAMs as the modelled system may not exhibit structural constancy between past and future.[5] Policy decisions informed by IAM analysis may even lead to changes in the causal relationships enshrined in a model's representation of energy, land use, and economic systems (DeCarolis et al. 2012; Weyant 2009). As an early historical example, the Club of Rome's 'Limits to Growth' scenarios from the 1970s are considered widely influential in shaping changes to resource management decision-making and policies (Nye 2004).

Second, IAMs are commonly used to define least-cost mitigation pathways to serve as normative reference points for what-if problems such as how to limit warming to 2 °C. Moreover, some IAMs are designed to find solutions which are inter-temporally optimal assuming perfect foresight over a 100-year timeframe. These normative applications of IAMs are not designed to reproduce how the modelled system actually behaves (Keppo et al. 2021). IAMs may also include optimisation elements to capture price formation in markets. However,

[5] Hodges and Dewar (1992) set out various conditions for testing model behaviour against observations, one of which is structural constancy. DeCarolis et al. (2012) give examples of how this condition may be violated in IAMs: 'Condition 2 requires that the 'causal structure' of the system being modeled remain constant through time. Energy economies at different geographic scales ... violate this condition. National priorities, technological change, and resource availability can result in structural economic shifts that are not captured by [energy-economy optimisation] models'. Note that the term 'structural change' is used differently in economic modelling to refer to changes in the sectoral composition of the economy and its value-adding activity. This type of structural change should be endogenously simulated, and does not represent a challenge to model evaluation.

real markets are imperfect and IAMs may not capture the numerous distortions through which observed prices are reflected (Trutnevyte 2016).

Third, IAMs focus on system responses to policy interventions relative to a dynamic and uncertain baseline, rather than an equilibrium (Rosen and Guenther 2016). As IAM baselines are dynamic, it is difficult to clearly separate drivers of change (e.g. economic growth, prices) from system responses (e.g. energy resource use, technology deployment, and greenhouse gas emissions).

These limitations are compounded by the practical challenge of finding observational data to describe historical energy and land-use systems in sufficient detail (Chaturvedi et al. 2013). Data challenges are more formidable in developing countries (van Ruijven et al. 2011), and prior to the 1970s when few energy data were systematically collected (Macknick 2011).

Historical simulations are therefore limited in their ability to give confidence in IAMs' representation of modelled systems (De Carolis 2011). But they remain a useful evaluation method under certain conditions: observational data are available; external drivers of change are clearly identifiable; the structure of modelled system components is constant; and normative characteristics can be relaxed.

4.2 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*.

Baseline scenarios from the IPCC Special Report on Emission Scenarios (SRES) were projected by IAMs in the late 1990s (Nakicenovic et al. 2000). These have been tracked against actual socioeconomic developments and emissions (Manning et al. 2010; Raupach et al. 2007; van Vuuren and O'Neill 2006). Through the 2000s, emission trends tended to track the upper bound of *ex ante* projections across a range of baseline assumptions (Peters et al. 2012). One implication is that scenario studies may inadequately capture uncertainty ranges in key drivers of change (Schwanitz and Wierling 2016). However, a recent comparison of multiple IPCC scenarios against observed fossil CO_2 emissions as well as population, economic, and energy system drivers over the period 1990–2020 found that observations have tended to track middle-of-the-road IAM projections (Strandsbjerg Tristan Pedersen et al. 2021).

IAM projections of energy prices and demand have also been compared against unfolding near-term observations (Craig et al. 2002; Pilavachi et al. 2008; Smil 2000). One recent study found near-term IAM projections underestimated the contribution of energy demand reductions to sustained CO_2 emission declines in 18 industrialised countries (Le Quéré et al. 2019). Near-term projections by IAMs of renewable energy technology deployment have also been found to consistently under-estimate observed growth rates, even under assumptions of stringent climate policy to limit warming to 2 °C (Creutzig et al. 2017).

Divergence from near-term observations is a potential source of insight for improving modelling efforts—if modellers look back at past projections (Koomey et al. 2003). However, modelled responses to policy interventions in the near term are not necessarily good indicators of long-term trends (van Vuuren et al. 2010).[6] IAMs are generally designed to represent long-

[6] Through the 2000s for example, IAM projections initially over-estimated emissions as a result of the Asia crisis and prolonged transition in Russia, then under-estimated emissions as the Chinese energy sector expanded rapidly with a heavy reliance on coal, then were brought back closer to emissions as a result of the financial crisis. Even *ex ante* projections such as the Lovins (1976) soft energy paths that closely matched near-term observations may have done so for the wrong reasons (Hafemeister 2007).

term dynamics such as the replacement of capital stock and path dependence from increasing returns to scale. Many IAMs also run on 5- or 10-year time steps which capture only multi-year averages (see Online Resources 5). However, IAMs have been applied recently to analyse *both* long-term *and* near-term outcomes of policy processes such as the national commitments made under the Paris Agreement (Vrontisi et al. 2018).

As with historical simulations, recent historical experience is useful for comparison against *ex ante* IAM projections only under certain conditions. First, only modelled processes with short-term characteristics or local responses should be tested against observations to ensure comparisons are clearly interpretable. Second, IAMs should resolve processes in short time steps (1–5 years) or have structural elements responsive to short-term drivers. Third, the system response to policy interventions (e.g. renewable energy regulation) or exogenous shocks (e.g. oil crises, collapse of the Soviet Union) should be clear and isolatable.

4.3 Stylised facts

An alternative method for drawing on history to evaluate IAMs examines whether generalisable historical patterns or 'stylised facts' are reproduced in model projections. This approach derives from the economist, Kaldor, who proposed 'a stylised view of the facts' that held when observing economic growth over long time periods, ignoring business cycles or other causes of volatility (Kaldor 1957; Leimbach et al. 2015).

The IPCC Special Report on Emission Scenarios (SRES) introduced comparisons of historical patterns in energy intensity and primary energy shares with future trends simulated by IAMs (Nakicenovic et al. 2000). Continuing in this vein, Schwanitz (2013) proposed a set of stylised facts describing aggregate long-term behavioural 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 (van Ruijven et al. 2008); durations of technology diffusion correlating positively with extents of diffusion (Wilson et al. 2012); and primary energy consumption correlating positively with economic growth (Schwanitz 2013). In each case, model projections were broadly consistent with historical dynamics, albeit with local or spatial differences. (For further details and examples, see Online Resources 3.)

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 technology deployment. Maximum projected rates of change are broadly consistent with maximum rates observed historically, even in scenarios with stringent climate policy (Iyer et al. 2015; van Sluisveld et al. 2015). Some studies have further triangulated between past, modelled futures, and expert opinions. Compared to IAM analysis, experts tend to have more bullish expectations for projected growth in renewable technologies, but more conservative expectations for fossil and nuclear technologies (van Sluisveld et al. 2018).

Testing the ability of IAMs to reproduce stylised facts is an additional way to draw on observational data to build confidence in structural representations of long-term system dynamics. However, the use of stylised facts as an evaluation method is restricted to aggregate system-level indicators or relationships, rather than specific causal mechanisms. This makes it hard to attribute any divergence from historical patterns.

4.4 Model hierarchies from simple to complex

Models such as IAMs face a tension between elaboration and elegance (Held 2005). More complex models may be more realistic, but may also be less tractable and interpretable: '… it is ironic that as we add more factors to a model, the certainty of its predictions may decrease even as our intuitive faith in the model increases' (Oreskes 2003). Whether a model has a 'good-enough' representation of the modelled system can be effectively tested through stripped-down versions designed to capture only the fundamental drivers of change (Jakeman et al. 2006).

'Model hierarchies' is a term used to describe models of the same system but spanning a range of complexity in terms of processes, dimensions, parameterisations, and spatial resolution (Held 2005; Stocker 2011). This is common in climate modelling: 'With the development of computer capacities, simpler models have not disappeared; on the contrary, a stronger emphasis has been given to the concept of a 'hierarchy of models' as the only way to provide a linkage between theoretical understanding and the complexity of realistic models' (p113, Treut et al. 2007). Climate models that are conceptually simpler or have less fine-grained resolution of processes and regions are useful for testing understanding of the modelled system. This helps interpret more complex models (Stainforth et al. 2007).[7]

In the early years of the Energy Modelling Forum (EMF), it was common practice to use simplified analytical frameworks to help interpret and understand larger-scale model results (Huntington et al. 1982; Thrall et al. 1983). This early practice of developing IAM model hierarchies has declined in more recent years. IAM development has tended inexorably towards ever finer scale resolution of ever more processes to assess ever more outcomes. For example, IAMs are now commonly used to analyse not just emission pathways but also progress towards a wide range of sustainable development goals (McCollum et al. 2018; von Stechow et al. 2015).[8]

One of the few fairly recent examples of a simple model being used to test structural uncertainty in a more complex IAM is the aptly named 'SIMPLE' model of global agriculture. This was designed to represent a minimal set of biophysical and economic relationships while still capturing the main drivers of global cropland use (Baldos and Hertel 2013). As an evaluation exercise, the SIMPLE model was tested to see if it could reproduce observed global trends in key indicators including crop land area, production, yield, and price over a historical simulation period from 1961 to 2006. The good fit to observations gave confidence in the basic model conceptualisation of land-use change dynamics that is embedded in more complex IAMs (Baldos and Hertel 2013).

There are many other opportunities for simple models of resource use, energy service demands, energy commodity trade, or technology deployment, to sit alongside complex global IAMs in model hierarchies which balance the competing merits of both elegance and elaboration. As well as enabling transparent testing of key elements of model structure, model

[7] Stainforth et al. (2007) go on to explain: 'It seems clear that the use of large nonlinear models is necessary in climate science but in the analysis of their output we must clearly identify assumptions which imply simpler linear models would have sufficed, at least until we understand the physics of the linear relations our complex models have revealed to us.'

[8] IAMs do have some simplified components including reduced-form climate or land-use models (Calvin and Fisher-Vanden 2017; Moore et al. 2017). These simplified model components can be tested against standalone models or sectoral models of greater complexity (van Vuuren et al. 2011).

hierarchies from simple to complex also mean appropriate models can be matched to the needs defined by particular research questions.

4.5 Model inter-comparisons

Model inter-comparison projects (MIPs) compare outputs, insights, and fits to observations across an ensemble of models. Like model hierarchies, MIPs are used to explore structural uncertainties in different models' representations of the same system.

Comparing results between multiple IAMs is a longstanding feature of climate mitigation analysis (Gaskins Jr. and Weyant 1993; Smith et al. 2015). The Energy Modelling Forum (EMF) started doing model comparisons in 1976, and in its early studies alternated policy-relevant MIPs with MIPs that were more diagnostic of model behaviour (Huntington et al. 1982; Sweeney and Weyant 1979). MIPs coordinated by EMF have contributed to IPCC assessments since 1995. (For further details including a historical timeline, see Online Resources 4.)

To enable comparability, MIPs require carefully designed experiments that harmonise key scenario assumptions (including external drivers and constraints) and standardise the reporting of model output (Huntington et al. 1982). IAM MIPs use controlled variations of policy assumptions (Clarke et al. 2009; Tavoni et al. 2015; Wilkerson et al. 2015), technology assumptions (Bosetti et al. 2015; Riahi et al. 2015), or socioeconomic development assumptions (Kriegler et al. 2016) or explore ensemble uncertainties. Other MIPs have focused on specific regions (van der Zwaan et al. 2016) or economic sectors (Ruane et al. 2017).

MIPs are a prominent evaluation method for IAMs, generating strong tacit learning for participating modelling teams. Within-ensemble agreement in IAM MIPs is often interpreted as providing 'robust' insights. However, agreement within the ensemble should be interpreted cautiously if structural differences between models are not systematic and models share approaches or components (Parker 2013). Policy decision-making informed by IAM MIPs should be based on 'estimates from many plausible, structurally distinct models' (Millner and McDermott 2016).

Model diagnostics are a specialised application of MIPs using a standardised set of indicators or performance metrics. These indicators classify model behaviour under harmonised scenario assumptions (Bennett et al. 2013). Diagnostic indicators therefore serve to 'fingerprint' models. Although descriptive, these indicators are an enabling step towards explaining characteristic IAM performance in terms of model structure and assumptions (Wilkerson et al. 2015). Model fingerprints also enable specific IAMs to be selected to match the analytical needs of specific scientific or policy questions. (For further details of the IAMs participating in these diagnostic studies, see Online Resources 5.)

4.6 Sensitivity analysis

Sensitivity analysis is used to identify model inputs and parameterisations influential on model output, and to attribute some of the uncertainties in outputs to uncertainties in inputs. By focusing on parametric uncertainty, sensitivity analysis is useful for testing the stability of the model over possible parameter ranges (e.g. to identify non-linearities or threshold effects in model behaviour). In IAMs, sensitivity analysis also helps identify value-laden parameters—like discount rates—that are influential on model outcomes (Beck and Krueger 2016). This guides further empirical research if uncertainties are parametric, or user-led appraisals of appropriate input ranges if uncertainties are societal (van der Sluijs et al. 2005).

Local or ‘one-at-a-time’ methods test output sensitivities to changes in single inputs or parameters; global methods vary multiple inputs or parameters simultaneously (Saltelli et al. 2008). Local sensitivities in IAMs are commonly reported in model evaluation studies. Influential inputs or parameters include rates of time preference (Belanger et al. 1993), rates of technological change (Sathaye and Shukla 2013), and investment costs of energy supply technologies (Koelbl et al. 2014). However, local sensitivity analyses on discrete parameters provide limited insights on how well a model represents the modelled system (Saltelli and D’Hombres 2010).

Global sensitivity analyses are also possible in IAMs using computationally efficient techniques (Borgonovo 2010). These have been used to explore the multi-dimensional global space spanned by uncertain model inputs and parameters, both in single models (Pye et al. 2015), and in MIPs (Bosetti et al. 2015; Marangoni et al. 2017; McJeon et al. 2011). This is a promising avenue of research for opening up the interpretability of IAM results in terms of input assumptions, particularly if reported alongside model applications (Mundaca et al. 2010). (For further details and examples, see Online Resources 6.)

However, care must be taken not to over-interpret results as model sensitivities may not correspond to those in real-world systems (Thompson and Smith 2019). The range over which inputs or parameters are sensitised may be informed by historical variation or defined arbitrarily (e.g. ± 10%). Formal techniques combining quantitative assessment with qualitative (expert) judgement can help inform both realistic ranges over which to vary parameters, and the interpretation of the sensitivity analysis (van der Sluijs et al. 2005).

4.7 Summary of strengths and limitations

Each of the IAM evaluation methods reviewed has certain restrictions on its application, and limitations on what can be learnt about model structure and behaviour. Each IAM evaluation method also faces conceptual, methodological, or practical challenges for its use and further development. Table 1 summarises the strengths and limitations of each method (see Online Resources 8 for further detail). Applying multiple evaluation methods in concert allows the limitations of one to be compensated by the strengths of another.

5 A systematic multi-method approach for strengthening IAM evaluation

Each evaluation method contributes to the testing of IAMs against one or more of the four evaluation criteria: appropriateness, interpretability, credibility, and relevance. These connections between method and criteria are shown as coloured arrows in the upper panel of Fig. 2. In the text below we explain each connection.

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 characterising general system dynamics. IAMs with specific causal mechanisms tested against observations in historical simulations or as stylised facts are appropriate for policy analysis linked to those

Table 1 Strengths and limitations of IAM evaluation methods

Evaluation method	Strengths	Limitations
Historical simulations	- Use of observations - Direct test of model behaviour and opportunity for quantitative comparison with observations	- Not as relevant for normative reference points (e.g. least-cost 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) - Computational cost for some models
Near-term observations	- Use of recent observations - Direct test of model behaviour in response to near-term policy forcings - Intuitive for users	- Models not designed to capture short-term variability - (*Other limitations as for historical simulations*)
Stylised facts	- Use of patterns in observations - Broad applicability (less sensitive to specific historical conditions and data constraints) - Test of model behaviour	- Subjective comparisons, no standardised tests - Difficulty of identifying reasons for divergence and implications for model structure
Model hierarchies from simple to complex models	- Tests understanding of key system processes - Links model behaviour to structural characteristics - Clearly interpretable results	- Limited applicability and policy relevance - Difficulty in simplifying heterogeneous causal mechanisms
Model inter-comparison projects (MIPs) inc. diagnostic indicators	- Identify results robust to (and sensitive to) structural uncertainty - Link differences between models' structure and parameterisation to differences in models' behaviour - Generalizable model classification or 'fingerprint' from standardised performance metrics - Peer review, exchange of data and methods, tacit learning among modelling teams	- Limited ability to implement harmonised scenario assumptions across models with different designs - Risk of over-tuning to harmonise diagnostic model runs - Difficulty of attributing divergent results to individual model differences - Risk of groupthink in shared modelling strategies, removal of outliers
Sensitivity analysis	- Identifies influential inputs and assumptions - Links model inputs and parameterisation to model behaviour	- Does not address structural uncertainty - Limited insights from local methods - Computational cost of global methods - Arbitrary selection of sensitivity ranges for inputs or parameters

mechanisms. These connections from evaluation method to the *appropriateness* criterion are shown as red arrows in Fig. 2.

IAMs vary widely in their resolution of mitigation measures, policy options, and spatial scales (Table 2.5M6 in Forster et al. 2018). Evaluating IAMs against the *appropriateness* criterion helps establish the 'resolution adequacy' of a model for a given task. Several evaluation methods contribute insights on resolution adequacy. For example, near-term observations that diverge from historical model projections may reveal influential processes omitted from model representations. Model intercomparison projects (MIPs) or diagnostic experiments involving models with different technological, process, and spatial resolutions are

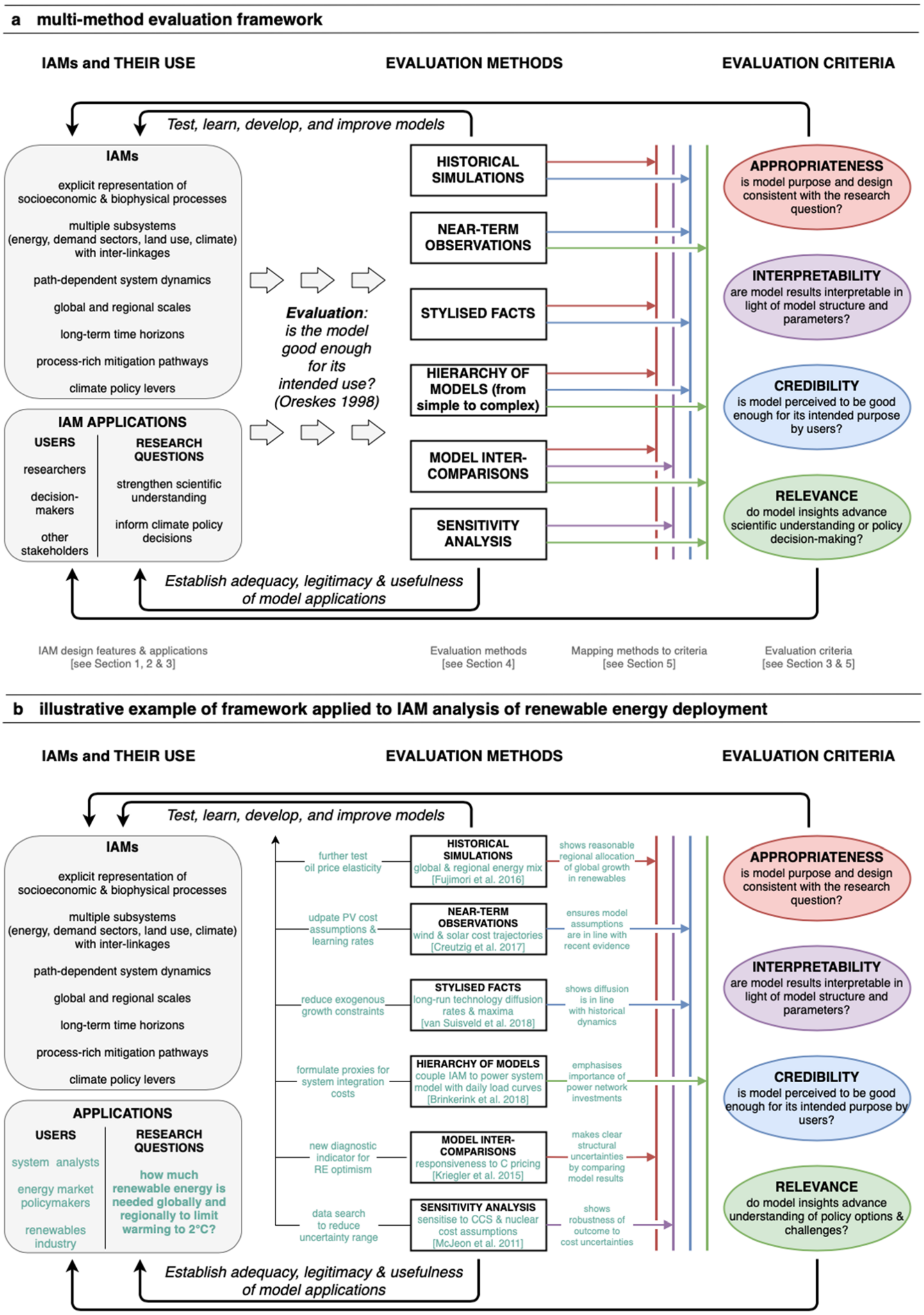

a multi-method evaluation framework
IAMs and THEIR USE
EVALUATION METHODS
EVALUATION CRITERIA
Test, learn, develop, and improve models
IAMs
explicit representation of socioeconomic & biophysical processes
multiple subsystems (energy, demand sectors, land use, climate) with inter-linkages
path-dependent system dynamics
global and regional scales
long-term time horizons
process-rich mitigation pathways
climate policy levers
IAM APPLICATIONS
USERS
researchers
decision-makers
other stakeholders
RESEARCH QUESTIONS
strengthen scientific understanding
inform climate policy decisions
Evaluation: is the model good enough for its intended use? (Oreskes 1998)
HISTORICAL SIMULATIONS
NEAR-TERM OBSERVATIONS
STYLISED FACTS
HIERARCHY OF MODELS (from simple to complex)
MODEL INTER-COMPARISONS
SENSITIVITY ANALYSIS
APPROPRIATENESS is model purpose and design consistent with the research question?
INTERPRETABILITY are model results interpretable in light of model structure and parameters?
CREDIBILITY is model perceived to be good enough for its intended purpose by users?
RELEVANCE do model insights advance scientific understanding or policy decision-making?
Establish adequacy, legitimacy & usefulness of model applications
IAM design features & applications [see Section 1, 2 & 3]
Evaluation methods [see Section 4]
Mapping methods to criteria [see Section 5]
Evaluation criteria [see Section 3 & 5]
b illustrative example of framework applied to IAM analysis of renewable energy deployment
IAMs and THEIR USE
EVALUATION METHODS
EVALUATION CRITERIA
Test, learn, develop, and improve models
IAMs
explicit representation of socioeconomic & biophysical processes
multiple subsystems (energy, demand sectors, land use, climate) with inter-linkages
path-dependent system dynamics
global and regional scales
long-term time horizons
process-rich mitigation pathways
climate policy levers
APPLICATIONS
USERS
system analysts
energy market policymakers
renewables industry
RESEARCH QUESTIONS
how much renewable energy is needed globally and regionally to limit warming to 2°C?
further test oil price elasticity
HISTORICAL SIMULATIONS global & regional energy mix [Fujimori et al. 2016]
shows reasonable regional allocation of global growth in renewables
udpate PV cost assumptions & learning rates
NEAR-TERM OBSERVATIONS wind & solar cost trajectories [Creutzig et al. 2017]
ensures model assumptions are in line with recent evidence
reduce exogenous growth constraints
STYLISED FACTS long-run technology diffusion rates & maxima [van Suisveld et al. 2018]
shows diffusion is in line with historical dynamics
formulate proxies for system integration costs
HIERARCHY OF MODELS couple IAM to power system model with daily load curves [Brinkerink et al. 2018]
emphasises importance of power network investments
new diagnostic indicator for RE optimism
MODEL INTER-COMPARISONS responsiveness to C pricing [Kriegler et al. 2015]
makes clear structural uncertainties by comparing model results
data search to reduce uncertainty range
SENSITIVITY ANALYSIS sensitise to CCS & nuclear cost assumptions [McJeon et al. 2011]
shows robustness of outcome to cost uncertainties
APPROPRIATENESS is model purpose and design consistent with the research question?
INTERPRETABILITY are model results interpretable in light of model structure and parameters?
CREDIBILITY is model perceived to be good enough for its intended purpose by users?
RELEVANCE do model insights advance understanding of policy options & challenges?
Establish adequacy, legitimacy & usefulness of model applications

Fig. 2 IAM evaluation methods test model adequacy against four evaluation criteria. Upper panel (a) shows general multi-method approach. Lower panel (b) shows illustrative example

useful for identifying the influence of resolution on the outcomes being explored in the MIP (see Online Resources 9 for further discussion).

MIPs also contribute to the *interpretability* of IAMs by linking model behaviour and resulting policy insights to structural representations of energy, land use, and economic processes. Sensitivity analysis similarly links model behaviour to input assumptions and parameter values. Standardisation of performance metrics and diagnostic indicators supports *interpretability* by 'fingerprinting' the distinctive behaviour of each model (Kriegler et al. 2015; Wilkerson et al. 2015). As IAMs represent multiple systems and their interactions, a pragmatic approach for improving *interpretability* is to evaluate individual model components sequentially (Harmsen et al. 2015; van Vuuren et al. 2011). These connections from evaluation methods to the *interpretability* criterion are shown as purple arrows in Fig. 2.

All three evaluation methods using historical data to test IAMs' abilities to reproduce observed short- and long-term dynamics are important for establishing *credibility*. In model hierarchies, simpler models have the further advantage of being more accessible to third parties (Craig et al. 2002; Crout et al. 2009; Edmonds and Reilly 1983). More transparent, documented, and accessible models and model results opening them up to independent review by a potentially diverse range of modellers, domain experts, and policy users (DeCarolis et al. 2017; DeCarolis et al. 2012; NCC 2015). This can include peer review or expert appraisal of models' theoretical consistency in conceptualising and representing the modelled system (Oppenheimer et al. 2016; van der Sluijs et al. 2005). These connections from evaluation methods to the *credibility* criterion are shown as blue arrows in Fig. 2.

Evaluation methods should similarly help strengthen the *relevance* of IAMs for advancing understanding of emission pathways and policy responses. Over the longer-term, both MIPs and sensitivity analyses help improve IAMs' ability to identify robust alternatives for achieving defined policy goals such as 2 °C climate stabilisation (De Carolis 2011; Drouet et al. 2015). Relevance for contemporary policy issues could be strengthened by testing IAM projections against near-term observations linked to clear policy levers, or by coupling IAMs to more detailed sectoral models to better capture constraints. These connections from evaluation methods to the *relevance* criterion are shown as green arrows in Fig. 2.

As Fig. 2 shows, the connections from methods to criteria are not unique. Multiple methods applied to single criteria may raise different issues or identify different grounds for improvement. This is entirely consistent with IAM evaluation as a continual process of learning and model development (Barlas 1996). A multi-method approach to IAM evaluation is necessary not just to ensure progress against the full set of evaluation criteria but also to compensate the limitations of some methods with the strengths of others (Table 1).

The lower panel of Fig. 2 gives an example of how this systematic IAM evaluation framework could be applied. We use renewable energy deployment for meeting climate targets as a specific IAM application. Whereas stylised facts may show long-term IAM projections conservatively in line with observed historical dynamics, near-term observations may reveal IAM underestimation of recent deployment trends. Resulting insights for IAM improvement include *both* relaxing long-term growth constraints *and* updating near-term technology cost and performance assumptions. Both insights from different methods strengthen the IAM's usefulness for the specific application. (For further details of this example, see Online Resources 8.)

The need for a systematic multi-method approach to IAM evaluation defines a major medium-to-long-term programme for IAM evaluation research and practice. Community-wide activities can play an important role in both driving and coordinating such a programme. High-value activities realisable in the near-term include:

- *Automating the calculation of diagnostic indicators in scenario databases* (Kriegler et al. 2015): to lower the threshold for diagnostic fingerprinting of new generations of IAMs.
- *Establishing community protocols for global sensitivity analyses over a wide range of variables for both individual modelling groups and as part of MIPs* (Marangoni et al. 2017): to increase the frequency, feasibility, and comparability of uncertainty analysis within the IAM community, and to prioritise evaluation research efforts for reducing parametric uncertainties.
- *Publishing community libraries of relevant historical data for hindcasting and stylised-facts experiments* (Schwanitz 2013): also helps to minimise the variation in MIPs due to unnecessary base-year differences.
- *Strengthening the use of empirical social science to develop and improve model representations of social and behavioural processes* (McCollum et al. 2017): to evaluate the influence of human decision-making and behavioural characteristics on important model outcomes.

Further examples of the effort required to standardise, interpret, facilitate, and best use IAM evaluation methods are provided in Online Resources 9.

6 Conclusion

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 synthesised many examples, benefits, insights, and limitations of applying different evaluation methods to IAMs. With the growing prominence of global IAMs in international and national climate policy, the time is ripe for establishing a more systematic approach to IAM evaluation, combining multiple methods in an ongoing, collaborative process involving both modellers and users.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10584-021-03099-9.

Author contribution CW wrote the manuscript with contributions from CG, EK, BvR, and DvV. All authors helped develop the research and commented on the manuscript.

Funding CW, CG, EK, DvV, and VK acknowledge funding support from EU FP7 grant #308329 (ADVANCE) and EU FP7 grant #265139 (AMPERE). ET acknowledges funding support from EPSRC grant #EP/P016847/1 (CRUISSE Network).

Availability of data and material Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Anderson K, Peters G (2016) The trouble with negative emissions. Science 354:182

Anthoff D, Tol RSJ (2013) The uncertainty about the social cost of carbon: a decomposition analysis using fund. Clim Chang 117:515–530

Baldos ULC, Hertel TW (2013) Looking back to move forward on model validation: insights from a global model of agricultural land use. Environ Res Lett 8:034024

Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12:183–210

Barlas Y, Carpenter S (1990) Philosophical roots of model validation: two paradigms. Syst Dyn Rev 6:148–166

Beck M, Krueger T (2016) The epistemic, ethical, and political dimensions of uncertainty in integrated assessment modeling. Wiley Interdiscip Rev Clim Chang 7(5):627–645 n/a-n/a

BEIS (2018) Updated short-term traded carbon values used for UK public policy appraisal. Department for Business, Energy and Industrial Strategy (BEIS), London

Belanger S, Cohan D, Deiner A, Drozd JM, Gjerde A, Peterson E (1993) The transition to reduced levels of carbon emissions. Energy Modeling Forum, Stanford

Bennett ND et al (2013) Characterising performance of environmental models. Environ Model Softw 40:1–20

Beugin D, Jaccard M (2012) Statistical simulation to estimate uncertain behavioral parameters of hybrid energy-economy models. Environ Model Assess 17:77–90

Borgonovo E (2010) Sensitivity analysis with finite changes: an application to modified EOQ models. Eur J Oper Res 200:127–138

Bosetti V et al (2015) Sensitivity to energy technology costs: a multi-model comparison analysis. Energy Policy 80:244–263

Cai Y, Judd KL, Lenton TM, Lontzek TS, Narita D (2015) Environmental tipping points significantly affect the cost−benefit assessment of climate policies. Proc Natl Acad Sci 112:4606–4611

Calvin K, Fisher-Vanden K (2017) Quantifying the indirect impacts of climate on agriculture: an inter-method comparison. Environ Res Lett 12:115004

Cash DW et al (2003) Knowledge systems for sustainable development. Proc Natl Acad Sci 100:8086

Chaturvedi V, Kim S, Smith SJ, Clarke L, Yuyu Z, Kyle P, Patel P (2013) Model evaluation and hindcasting: an experiment with an integrated assessment model. Energy 61:479–490

Clarke L, Edmonds J, Krey V, Richels R, Rose S, Tavoni M (2009) International climate policy architectures: overview of the EMF 22. Int Scenarios Energy Econ 31:S64–S81

Clarke L et al (2014) Chapter 6: Assessing transformation pathways. In: Working Group III contribution to the IPCC 5th Assessment Report, Climate Change 2014: Mitigation of climate change. Cambridge University Press, Cambridge

Cooke RM (2015) Messaging climate change uncertainty Nature Clim Chang 5:8–10

Craig PP, Gadgil A, Koomey JG (2002) What can history teach us? A retrospective examination of long-term energy forecasts for the United States. Annu Rev Energy Environ 27:83–118

Creutzig F, Agoston P, Goldschmidt JC, Luderer G, Nemet G, Pietzcker RC (2017) The underestimated potential of solar energy to mitigate climate change. Nat Energy 2:17140

Crout NMJ, Tarsitano D, Wood AT (2009) Is my model too complex? Evaluating model formulation using model reduction. Environ Model Softw 24:1–7

DeCarolis J et al (2017) Formalizing best practice for energy system optimization modelling. Appl Energy 194: 184–198

De Carolis JF (2011) Using modeling to generate alternatives (MGA) to expand our thinking on energy futures. Energy Econ 33:145–152

DeCarolis JF, Hunter K, Sreepathi S (2012) The case for repeatable analysis with energy economy optimization models. Energy Econ 34:1845–1853

Drouet L, Bosetti V, Tavoni M (2015) Selection of climate policies under the uncertainties in the Fifth Assessment Report of the IPCC. Nat Clim Chang 5:937–940

Edenhofer O, Minx J (2014) Mapmakers and navigators, facts and values. Science 345:37–38

Edmonds J, Reilly J (1983) A long-term global energy- economic model of carbon dioxide release from fossil fuel use. Energy Econ 5:74–88

Eker S, Rovenskaya E, Obersteiner M, Langan S (2018) Practice and perspectives in the validation of resource management models. Nat Commun 9:5359

Fischhoff B (2015) The realities of risk-cost-benefit analysis. Science 350:aaa6516

Flato G et al (2013) Evaluation of climate models. In: Stocker TF et al (eds) Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge

Forster P, Huppmann D, Kriegler E, Mundaca L, Smith C, Rogelj J, Séférian R (2018) Mitigation pathways compatible with 1.5°C in the context of sustainable development: supplementary material. In: Masson-Delmotte V et al (eds) Global warming of 1.5°C: an IPCC special report. World Meteorological Organization, Geneva, pp 93–174

Fujimori S, Dai H, Masui T, Matsuoka Y (2016) Global energy model hindcasting. Energy 114:293–301

Gargiulo M, Gallachóir BÓ (2013) Long-term energy models: principles, characteristics, focus, and limitations. Wiley Interdiscip Rev: Energy and Environ 2:158–177

Gaskins DW Jr, Weyant JP (1993) Model comparisons of the costs of reducing CO2 emissions. Am Econ Rev 83:318–323

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

Grassi G et al (2018) Reconciling global-model estimates and country reporting of anthropogenic forest CO2 sinks. Nature Climate Change 8:914–920

Greenstone M, Kopits E, Wolverton A (2013) Developing a social cost of carbon for US regulatory analysis: a methodology and interpretation. Rev Environ Econ Policy 7:23–46

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 37:4259–4266

Hafemeister D (2007) Physics of societal issues: calculations on national security, environment, and energy. Springer, Berlin

Harmsen MJHM et al (2015) How well do integrated assessment models represent non-CO2 radiative forcing? Clim Chang 133:565–582

Held IM (2005) The gap between simulation and understanding in climate modeling. Bull Am Meteorol Soc 86: 1609–1614

Hibino G, Pandey R, Matsuoka Y, Kainuma M (2013) A guide to AIM-enduse model. National Institute of Environmental Studies (NIES), Tsukuba

Hodges JS, Dewar JA (1992) Is it you or your model talking? A framework for model validation. RAND Corporation

Hope C, Hope M (2013) The social cost of CO2 in a low-growth world. Nat Clim Chang 3:722–724

Huntington HG, Weyant JP, Sweeney JL (1982) Modeling for insights, not numbers: the experiences of the energy modeling forum. Omega 10:449–462

Iyer G, Hultman N, Eom J, McJeon H, Patel P, Clarke L (2015) Diffusion of low-carbon technologies and the feasibility of long-term climate targets. Technol Forecast Soc Chang 90:103–118

Jakeman AJ, Letcher RA, Norton JP (2006) Ten iterative steps in development and evaluation of environmental models. Environ Model Softw 21:602–614

Kaldor N (1957) A model of economic growth. Econ J 67:591–624

Keppo I et al. (2021) Exploring the possibility space: taking stock of the diverse capabilities and gaps in integrated assessment models. Environ Res Lett

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

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. Int J Greenhouse Gas Control 27:81–102

Koomey J, Craig P, Gadgil A, Lorenzetti D (2003) Improving long-range energy modeling: a plea for historical retrospectives. Energy J 24:75–92

Krey V et al (2016) MESSAGE-GLOBIOM 1.0 Documentation. International Institute for Applied Systems Analysis (IIASA), Laxenburg

Krey V et al (2014) Annex II: Metrics & methodology. In: Working Group III contribution to the IPCC 5th Assessment Report, Climate Change 2014: Mitigation of Climate Change. Cambridge University Press, Cambridge

Kriegler E et al (2016) Will economic growth and fossil fuel scarcity help or hinder climate stabilization? Clim Chang 136:7–22

Kriegler E et al (2015) Diagnostic indicators for integrated assessment models of climate policy. Technol Forecast Soc Chang 90:45–61

Kunreuther H et al (2014) Integrated risk and uncertainty assessment of climate change response policies. In: Edenhofer O et al (eds) Working Group III contribution to the IPCC 5th Assessment Report, Climate Change 2014: Mitigation of Climate Change. Cambridge University Press, Cambridge

Le Quéré C et al (2019) Drivers of declining CO2 emissions in 18 developed economies. Nat Clim Chang 9:213–217

Leimbach M, Kriegler E, Roming N, Schwanitz J (2015) Future growth patterns of world regions – a GDP scenario approach. Glob Environ Chang 42:215–225

Lovins A (1976) Energy strategy: the road not taken? Foreign Affairs 55:65–96

Luderer G, Pietzcker R, Bertram C, Kriegler E, Meinshausen M, Edenhofer O (2013) Economic mitigation challenges: how further delay closes the door for achieving climate targets. Environ Res Lett 8:034033

Macknick J (2011) Energy and CO2 emission data uncertainties. Carbon Manag 2:189–205

Manne AS, Richels RG (1992) Estimating the energy conservation parameters: an experiment in backcasting. In: Manne AS, Richels RG (eds) Buying greenhouse insurance: the economic costs of carbon dioxide emission limits. MIT press, Cambridge

Manning MR et al (2010) Misrepresentation of the IPCC CO2 emission scenarios. Nat Geosci 3:376–377

Marangoni G et al (2017) Sensitivity of projected long-term CO2 emissions across the Shared Socioeconomic Pathways. Nat Clim Chang 7:113–117

McCollum DL et al (2018) Connecting the sustainable development goals by their energy inter-linkages. Environ Res Lett 13:033006

McCollum DL et al (2017) Improving the behavioral realism of global integrated assessment models: an application to consumers' vehicle choices. Transp Res Part D: Transp Environ 55:322–342

McDowall W, Trutnevyte E, Tomei J, Keppo I (2014) Reflecting on Scenarios. UKERC (UK Energy Research Centre) Energy Systems Theme, London

McJeon HC, Clarke L, Kyle P, Wise M, Hackbarth A, Bryant BP, Lempert RJ (2011) Technology interactions among low-carbon energy technologies: what can we learn from a large number of scenarios? Energy Econ 33:619–631

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

Millner A, McDermott TKJ (2016) Model confirmation in climate economics. Proc Natl Acad Sci 113:8675–8680

Moore FC, Baldos ULC, Hertel T (2017) Economic impacts of climate change on agriculture: a comparison of process-based and statistical yield models. Environ Res Lett 12:065008

Moore FC, Diaz DB (2015) Temperature impacts on economic growth warrant stringent mitigation policy. Nat Clim Chang 5:127–131

Moss RH et al (2010) The next generation of scenarios for climate change research and assessment. Nature 463: 747–756

Mundaca L, Neij L, Worrell E, McNeil M (2010) Evaluating energy efficiency policies with energy-economy models. Annu Rev Environ Resour 35:305–344

Nakicenovic N et al (2000) Special report on emissions scenarios. Cambridge University Press, Cambridge

Nakicenovic N, Lempert RJ, Janetos AC (2014) A framework for the development of new socio-economic scenarios for climate change research: introductory essay. Clim Chang 122:351–361

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

NCC (2015) IAM helpful or not? Nat Clim Chang 5:81–81

Nordhaus WD (2013) The climate casino: risk, uncertainty, and economics for a warming world. Yale University Press, New Haven

Nye DE (2004) Technological prediction: a promethean problem. In: Sturken M, Thomas D (eds) Technological visions: the hopes and fears that shape new technologies. Temple University Press, Philadelphia, pp 159–176

Oppenheimer M, Little CM, Cooke RM (2016) Expert judgement and uncertainty quantification for climate change. Nat Clim Chang 6:445–451

Oreskes N (1998) Evaluation (not validation) of quantitative models. Environ Health Perspect 106:1453–1460

Oreskes N (2003) The role of quantitative models in science. In: Canham CD, Cole JJ, Lauenroth WK (eds) Models in ecosystem science. Princeton University Press, Princeton, pp 13–31

Oreskes N, Shrader-Frechette K, Belitz K (1994) Verification, validation, and confirmation of numerical models in the earth sciences. Science 263:641–646

Parker WS (2013) Ensemble modeling, uncertainty and robust predictions. Wiley Interdiscip Rev Clim Chang 4: 213–223

Peace J, Weyant J (2008) Insights not numbers: the appropriate use of economic models. Pew center on global climate change, Washington

Peters GP et al (2012) The challenge to keep global warming below 2oC. Nat Clim Chang 3:4–6

Pilavachi PA, Dalamaga T, Rossetti di Valdalbero D, Guilmot JF (2008) Ex-post evaluation of European energy models. Energy Policy 36:1726–1735

Pirtle Z, Meyer R, Hamilton A (2010) What does it mean when climate models agree? A case for assessing independence among general circulation models. Environ Sci Pol 13:351–361

Pye S, Sabio N, Strachan N (2015) An integrated systematic analysis of uncertainties in UK energy transition pathways. Energy Policy 87:673–684

Raupach ME, Marland G, Ciais P, Quéré CL, Canadell JG, Klepper G, Field CB (2007) Global and regional drivers of accelerating CO2 emissions. Proc Natl Acad Sci 104:10288–10293

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

Riahi K et al (2017) The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: an overview. Glob Environ Chang 42:153–168

Rogelj J et al (2018) Mitigation pathways compatible with 1.5°C in the context of sustainable development. In: Masson-Delmotte V et al (eds) Global warming of 1.5°C: an IPCC special report. World Meteorological Organization, Geneva, pp 93–174

Rosen RA, Guenther E (2015) The economics of mitigating climate change: what can we know? Technol Forecast Soc Chang 91:93–106

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

Ruane AC et al (2017) An AgMIP framework for improved agricultural representation in integrated assessment models. Environ Res Lett 12:125003

Saltelli A, D'Hombres B (2010) Sensitivity analysis didn't help. A practitioner's critique of the Stern review. Glob Environ Chang 20:298–302

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Tarantola S (2008) Global sensitivity analysis: the primer. Wiley, Chichester

Sargent RG (2013) Verification and validation of simulation models. J Sim 7:12–24

Sathaye J, Shukla PR (2013) Methods and models for costing carbon mitigation. Annu Rev Environ Resour 38: 137–168

Schindler DE, Hilborn R (2015) Prediction, precaution, and policy under global change. Science 347:953–954

Schneider SH (1997) Integrated assessment modeling of global climate change: transparent rational tool for policy making or opaque screen hiding valueladen assumptions? Environ Model Assess 2:229–249

Schwanitz VJ (2013) Evaluating integrated assessment models of global climate change. Environ Model Softw 50:120–131

Schwanitz VJ, Wierling A (2016) Offshore wind investments – realism about cost developments is necessary. Energy 106:170–181

Smil V (2000) Perils of long-range energy forecasting: reflections on looking far ahead. Technol Forecast Soc Chang 65:251–264

Smith SJ et al (2015) Long history of IAM comparisons. Nat Clim Chang 5:391

Snyder AC, Link RP, Calvin KV (2017) Evaluation of integrated assessment model hindcast experiments: a case study of the GCAM 3.0 land use module. Geosci Model Dev 10:4307–4319

Stainforth DA, Allen MR, Tredger ER, Smith LA (2007) Confidence, uncertainty and decision-support relevance in climate predictions. Philos Trans R Soc London A: Math, Phys Eng Sci 365:2145–2161

Stern N (2006) The Stern review on the economics of climate change. Cambridge University Press, Cambridge

Stern N (2016) Current climate models are grossly misleading. Nature 530:407–409

Stocker T (2011) Model hierarchy and simplified climate models. In: In: Introduction to climate modelling. Advances in geophysical and environmental mechanics and mathematics. Springer, Berlin Heidelberg, pp 25–51

Strandsbjerg Tristan Pedersen J, Duarte Santos F, van Vuuren D, Gupta J, Encarnação Coelho R, Aparício BA, Swart R (2021) An assessment of the performance of scenarios against historical global emissions for IPCC reports. Glob Environ Chang 66:102199

Sweeney JL, Weyant JP (1979) The energy modeling forum: past, present and future. In: Nemetz PN (ed) Energy policy: the global challenge. Institute for Research on Public Policy, Montreal, pp 295–320

Tavoni M et al (2015) Post-2020 climate agreements in the major economies assessed in the light of global models. Nat Clim Chang 5:119–126

Tebaldi C, Knutti R (2007) The use of the multi-model ensemble in probabilistic climate projections. Philos Trans R Soc A: Mathematical, Phys Eng Sci 365:2053–2075

Thompson EL, Smith LA (2019) Escape from model-land. Kiel Institute for the World Economy, Kiel

Thrall RM, Thompson RG, Holloway ML (1983) Large-scale energy models: prospects and potential. Westview Press, Boulder

Toth FL (1994) Practice and progress in integrated assessments of climate change: a review. In: Nakicenovic N, Nodhaus, W.D., Richels, R. and Toth, F.L. (ed) Integrative assessment of mitigation, impacts and adaptation to climate change, IIASA, Laxenburg, Austria., 13–15 October 1993

Treut HL et al (2007) Historical overview of climate change. In: Solomon S et al (eds) Climate change 2007: the physical science basis, I edn. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge

Trutnevyte E (2016) Does cost optimisation approximate the real-world energy transition? Energy 106:182–193

UNEP (2015) The Emissions Gap Report 2015. United Nations Environment Programme (UNEP), Nairobi

UNFCCC (2015) Synthesis report on the aggregate effect of the intended nationally determined contributions. United Nations Framework Convention on Climate Change (UNFCCC) Secretariat, Bonn

van Beek L, Hajer M, Pelzer P, van Vuuren D, Cassen C (2020) Anticipating futures through models: the rise of Integrated Assessment Modelling in the climate science-policy interface since 1970. Glob Environ Chang 65:102191

van der Sluijs JP, Craye M, Funtowicz S, Kloprogge P, Ravetz J, Risbey J (2005) Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: the NUSAP system. Risk Anal 25:481–492

van der Sluijs JP, Petersen AC, Janssen PHM, Risbey JS, Ravetz JR (2008) Exploring the quality of evidence for complex and contested policy decisions. Environ Res Lett 3:024008

van der Zwaan BCC, Calvin KV, Clarke LE (2016) Climate mitigation in Latin America: implications for energy and land use: preface to the Special Section on the findings of the CLIMACAP-LAMP project. Energy Econ 56:495–498

van Ruijven B, Sluijs J, Vuuren D, Janssen P, Heuberger PC, Vries B (2010) Uncertainty from model calibration: applying a new method to transport energy demand modelling. Environ Model Assess 15:175–188

van Ruijven B, Urban F, Benders RMJ, Moll HC, van der Sluijs JP, de Vries B, van Vuuren DP (2008) Modeling energy and development: an evaluation of models and concepts. World Dev 36:2801–2821

van Ruijven B, van Vuuren DP, de Vries B, Isaac M, van der Sluijs JP, Lucas PL, Balachandra P (2011) Model projections for household energy use in India. Energy Policy 39:7747–7761

van Sluisveld M et al (2015) Comparing future patterns of energy system change in 2°C scenarios with historically observed rates of change. Glob Environ Chang 35:436–449

van Sluisveld MAE, Harmsen MJHM, van Vuuren DP, Bosetti V, Wilson C, van der Zwaan B (2018) Comparing future patterns of energy system change in 2 °C scenarios to expert projections. Glob Environ Chang 50:201–211

van Vuuren DP et al (2010) What do near-term observations tell us about long-term developments in greenhouse gas emissions? Glob Environ Chang 103:635–642

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. Technol Forecast Soc Chang 98:303–323

van Vuuren DP et al (2011) How well do integrated assessment models simulate climate change? Clim Chang 104:255–285

van Vuuren DP, O'Neill BC (2006) The consistency of IPCC's SRES scenarios to 1990–2000 trends and recent projections. Clim Chang 75:9–46

von Stechow C et al (2015) Integrating global climate change mitigation goals with other sustainability objectives: a synthesis. Annu Rev Environ Resour 40:363–394

Vrontisi Z et al (2018) Enhancing global climate policy ambition towards a 1.5 °C stabilization: a short-term multi-model assessment. Environ Res Lett 13:044039

Waisman H, Guivarch C, Grazi F, Hourcade J-C (2011) The Imaclim-R model: infrastructures, technical inertia and the costs of low carbon futures under imperfect foresight. Clim Chang 114:101–120

Weitzel M et al (2019) Model-based assessments for long-term climate strategies. Nat Clim Chang 9:345–347

Weyant J (2009) A perspective on integrated assessment: an editorial comment. Clim Chang 95:317–323

Weyant J (2017) Some contributions of integrated assessment models of global climate change. Rev Environ Econ Policy 11:115–137

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

Wilson C, Grubler A, Bauer N, Krey V, Riahi K (2012) Future capacity growth of energy technologies: are scenarios consistent with historical evidence? Clim Chang 118:381–395

Wynne B (1984) The institutional context of science, models, and policy: the IIASA energy study. Policy Sci 17: 277–320

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Charlie Wilson[1,2] • **Céline Guivarch**[3] • **Elmar Kriegler**[4,5] • **Bas van Ruijven**[2] • **Detlef P. van Vuuren**[6,7] • **Volker Krey**[2] • **Valeria Jana Schwanitz**[8,9] • **Erica L. Thompson**[10]

[1] Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, Norwich, Norfolk NR4 7TJ, UK

[2] International Institute for Applied Systems Analysis (IIASA), Schlossplatz 1, 2361 Laxenburg, Austria

[3] Centre International de Recherche sur l’Environnement et le Developpement (CIRED), Ecole des Ponts ParisTech, 45 bis, Av. de la Belle Gabrielle, F-94736 Nogent-sur-Marne, France

[4] Potsdam Institute for Climate Impact Research, Member of The Leibniz Association, P.O. Box 601203, 14412 Potsdam, Germany

[5] Faculty of Economics and Social Sciences, University of Potsdam, August-Bebel-Str. 89, 14482 Potsdam, Germany

[6] PBL Netherlands Environmental Assessment Agency, P.O. Box 303, 3720 AH Bilthoven, The Netherlands

[7] Copernicus Institute of Sustainable Development, Utrecht University, Heidelberglaan 2, 3584 CS Utrecht, The Netherlands

[8] Western Norway University of Applied Sciences, postboks 7030, 5020 Bergen, Norway

[9] The Schumacher Institute, BS1 6XN Bristol, UK

[10] Data Science Institute, London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK