Evaluating integrated assessment models of global climate change

Valeria Jana Schwanitz

Potsdam Institute for Climate Impact Research, Telegraphenberg A 31, Postfach 60 12 03, D-14412 Potsdam

Abstract

Integrated Assessment Models of global climate change (IAMs) are an established tool to study interlinkages between the human and the natural system. Insights from these complex models are widely used to advise policy-makers and to inform the general public. But up to now there has been little understanding of how these models can be evaluated and community-wide standards are missing. To answer this urgent question is a challenge because the systems are open and their future behavior is fundamentally unknown. In this paper, we discuss ways to overcome these problems. Reflecting on experience from other modelling communities, we develop an evaluation framework for IAM of global climate change. It builds on a systematic and transparent step-by-step demonstration of a model's usefulness testing the plausibility of its behavior. Steps in the evaluation hierarchy are: setting up an evaluation framework, evaluation of the conceptual model, code verification and documentation, model evaluation, uncertainty and sensitivity analysis, documentation of the evaluation process, and communication with stakeholders. An important element in evaluating IAM of global climate change is the use of stylized behavior patterns derived from historical observation. The discussion of two examples is offered in this paper.

Keywords: model evaluation, model validation, integrated assessment models, transparency, community tests and standards, global climate change, stylized facts

Email address: schwanitz@pik-potsdam.de (Valeria Jana Schwanitz)

1. Introduction

Integrated assessment models (short: IAMs) of global climate change are an important tool to study human feedbacks and influences on climate change and mitigation of greenhouse gases. The advantage of these complex models is that they provide an integrated system perspective. For this purpose, different models are coupled with each other, e.g. a climate model, a land-use model, an energy model, and a model describing economic growth. The results of IAMs are being used for advising policy-makers and informing the global public. Recent examples of publications with contributions from IAMs are the Special Report on Renewable Energies (IPCC, 2011) or the World Energy Outlook 2011 (International Energy Agency, 2011). Based on the results of IAMs, long-term policy plans such as the EU Energy Roadmap 2050 (EC COM, 2011) are formulated; IAMs also provide background information for international climate policy negotiations. It is therefore very legitimate to ask how much one can trust these 'big' models to deliver reliable answers and to demand an appropriate communication of assumptions, references, errors, and uncertainty ranges. This is and has been very much the concern of the IAM community itself and an ultimate answer has not yet been found. Up to now there has not been a community-wide understanding of what validation of IAMs could mean nor are there standards or protocols on how to evaluate them (Risbey et al., 1996; Parker et al., 2002). The aim of this paper is to contribute to the emerging discussion in the community by proposing a definition of evaluation of IAMs as well as a framework for performing evaluation exercises. Such frameworks have been set-up in scientific engineering (e.g. Sargent (2003, 2010); Oberkampf and Roy (2010)) as well as in the broader environmental IAM community (e.g. Konikow and Bredehoeft (1992); Rykiel (1996); Schneider (1997); Jakeman et al. (2006)). Even the involvement of non-specialists has been advanced (Voinov and Bousquet, 2010; Krueger et al., 2012).

One could object to such additional efforts by pointing towards common quality standards in scientific publishing. Beck et al. (1997), however, state that neither common scientific peer-review processes nor history matching exercises can sufficiently solve the problem if models are continually growing in complexity, see also Konikow and Bredehoeft (1992). It becomes more and more difficult for third parties outside of the modelling teams to scrutinize their publications with reasonable effort. Given this problem on the one hand and the relevance of IAMs for the policy-making process on the other

hand, the central questions are the following: can IAMs be validated? How is it possible to build trust in their output and structure? How should IAM results be interpreted? What is a transparent and comprehensible way of communicating assumptions and uncertainties of these complex models?

Model validation is and has been a controversial issue because it entails consequences: what if a model's performance is poor? What are the (opportunity) costs of developing evaluation routines and adopting performance standards? This is a very relevant problem given that the development of IAMs is strongly driven by policy demand and third-party funded community projects. However, the strong ties between funding possibilities and applied research questions that satisfy the demand of a rapidly changing society leave too little room for work-intensive, formal model validation. Hence, it does not come as a surprise that validation is not a top priority even though the demand for it is high.

Up to now, we have implied that an agreement on how to validate IAMs can be reached and standards can be developed and adopted. This is a brave assumption as the discussion already gets vivid when it comes to the wording: is the validation of earth system models impossible from a strict philosophical viewpoint (Oreskes et al., 1994)? Or is it "corroboration" (Oderwald and Hans, 1993) or "confirmation" (Oreskes et al., 1994; Carolan, 2008)? Is "evaluation" the better choice, as it is preferred in the climate modelling community? Barlas and Carpenter (1990) point out that validation of system dynamic models "is inherently a social, judgmental qualitative process". Can the evaluation process be named "validation" because it is "validation" that the third parties are expecting? (It would be necessary of course to explain at the same time what can be expected from a validated IAM.) Such an approach would be very much in the spirit of Celia et al. (1992) who note that the "perception of validation" matters and not the "semantics". Others, however, would strictly reject this, see Konikow and Bredehoeft (1992). The authors state that "emphasizing validation deceives society with the impression that, by expending sufficient effort, uncertainty can be eliminated and absolute knowledge be attained." It is, however, not the intention of this paper to solve the semantic puzzle. We stay with the more neutral term 'evaluation' throughout this paper.

This paper states that evaluation of IAMs of global climate change should be understood as a continuous effort of testing whether the model can fulfill its purpose. Documentation of the model, including transparency about its shortcomings and area of applicability, are integral to the evaluation process.

Given the fundamental problem for IAMs which is the lack of real-system data, evaluation exercises should comprise a variety of tests. Moreover, the focus should be to assess the plausibility of a model's explanatory power as its forecasting power cannot be validated.

The paper is organized as follows: the next section starts with a brief review of challenges for evaluating IAMs, most of which are connected with the properties of open systems whose future behavior is fundamentally unknown. This a-priori lack of experimental data adds a further challenge to their evaluation in comparison to IAMs in the non-climate context. For these models, data for comparison sometimes exist or can at least be generated. The discussion is followed in Section 3 with an overview on how validation/evaluation is defined in different modelling communities and how different aspects relate to IAMs. The aim is to provide a theoretical background for the discussion of how the evaluation process could be practically organized (Section 4.). In this section different steps of the proposed evaluation hierarchy are illustrated along with examples. The final section summarizes the paper by framing action items to the IAM community of global climate change.

2. Challenges in evaluating IAMs of global climate change

2.1. General challenges of dynamic large scale models

In their reviews, Barlas and Carpenter (1990); Barlas (1996) argue that the paradigm of system dynamic modelling matches the "relativist/functional/holistic" philosophy of science as it is impossible to establish formal objectivity. The authors emphasize that, first, models are being constructed for specific purposes and any validation method has to be designed around the model purpose. Second, apart from their predictive capabilities, the explanatory power of system dynamic models is important. It is not only of interest whether the model output is right (i.e. matching with observations from the modelled system), but also that one sees the right results for the right reasons. The latter is referred to as "structure validity" in Barlas (1996), which precedes tests of the output behavior ("behavior validity").

However, conceptual models in system dynamics modelling are biased by disciplines as the agreement about the appropriate level of detail is not necessarily given. Economists might prefer a more detailed description of economic development, while energy engineers might emphasize a higher resolution of technological processes. This is important because objective criteria on what

is the correct model view cannot be defined. Third, model results are often not suitable for statistical testing as output variables are auto-correlated and cross-correlated. The absence of a single important output variable also raises the problem of multi-hypothesis testing. Finally, the authors raise the point that it is impossible to establish an objective significance level.

Beven (2002) starts from the recognition of a common "pragmatic realism" with the aim to model a system as realistic as possible without too much worries about a philosophical grounding. The author suggests instead to acknowledge that environmental models are never true. This leads to a pragmatic notion of a relativist understanding of the nature of knowledge, see also discussions following the paper (Beven, 2004). An important principle is that system events are neither unique nor knowable. Therefore, different "behavioral" representations of the system are possible (equifinality of models).

In a widely recognized paper, Jakeman et al. (2006) propose steps for developing and evaluating a larger class of environmental models. The authors point out that formal falsification and statistical hypothesis testing "is rarely possible (or perhaps even appropriate) for large, integrated models".

Regarding the evaluation of IAMs of global climate change, there are two more fundamental challenges: they describe systems which are open and whose future behavior is fundamentally unknown.

2.2. IAMs of global climate change are open systems

In a seminal paper on issues of verification and validation of earth system models, Oreskes et al. (1994) underline the philosophical argument that the truth of a proposition can only be proven (i.e. validated) if the system is closed[1]. There are several reasons why models can be incomplete (see also Risbey et al. (1996)). In the following, we discuss this issue in the specific context of IAMs.

The first reason is already obvious from a practical point of view: due to its mere complexity, it is impossible to perfectly mimic the universe and it becomes a matter of delimitation of the system. As Risbey et al. (1996) argue, this leads to a bias as one has to prioritize what should be integrated into the model and what should be omitted. Most often the decision is

[1]An open (i.e. not closed) system is a system where interactions between internal elements of the system and the system's environment may happen.

simply guided by practicability: one implements what is relatively easy to implement.

Second, fundamental laws and processes of the system are incompletely known. It must be remembered that IAMs are an attempt to integrate the ecosphere and the anthroposphere. In particular, the laws governing human behavior and decision making are largely Terra Incognita and related parameters are the subject of vivid debates, e.g. see the discussion following the Stern review (Stern, 2007; Nordhaus, 2007; Tol, 2006; Weitzman, 2007). The problem of incomplete knowledge is amplified by the fact that IAMs are exploring the space of possible futures. If we already do not fully understand historical laws and processes related to human behavior and preferences, how can we be confident about their future values - let alone their dynamic patterns? As matter-of-fact, preference parameters in IAMs of global climate change are often set as constant for a period of 100 years, e.g. discount rates do not change over time. Given fundamental system changes that occur at these time scales, this is a very unrealistic assumption. Another Terra Incognita is the creation and diffusion of knowledge. We simply do not know how the future will unfold. For example, can it be expected to have a back-stop technology available in due time for solving the climate problem? Therefore, IAM modellers seldom have no better choice but to use subjective rules of thumb and rough approximations. Apart from incomplete knowledge about fundamental laws of the system, model parameters are also incomplete as the uncertainty inherent in available data is high, accounting methods differ, and data series are incomplete for several reasons (see Macknick (2011) who discuss uncertainty in emission and energy data across various data sources).

Complexity of the system and incomplete knowledge about its fundamental laws give rise to variables and models that are laden with interferences and assumptions. This leads to the introduction of a bias (Risbey et al., 1996; Carolan, 2008; Rosenberg, 1994). A prominent example can be found in Risbey et al. (1996) summarizing approaches across disciplines on how to value human life. A modelling team's subjective opinion also enters an IAM when a choice has to be made on how to take the future development into account; modelling approaches range from inherently myopic to perfect foresight models.

Furthermore, many non-additive properties are being scaled-up in IAMs. This concerns fundamental, open questions on the invariance of scales. To give economic examples: what is the micro foundation of macro-economic

development? How are the non-homogeneous decisions of households, business enterprises, and governments best represented in a model with a global perspective? What technological detail is needed to capture relevant processes? To answer such questions, IAM modellers use different approaches. For example, there are bottom-up, top-down, and hybrid formulations of energy-economy models (see e.g. Hourcade et al. (2006) for a review). Furthermore, IAMs of global climate change divide the world into several macro regions whose input parameters are derived from sub-national or national data.

Finally, scales can differ across IAM components (Parker et al., 2002). For example, while important processes in the climate system are very slow, taking centuries or millennia, important processes in the economy or energy system are closer to the planning horizon of human beings. In addition, processes at lower scales are often being omitted, since the typical resolution of time in IAMs amounts to 1-5 years. Thus, processes that occur on a daily or monthly basis are ignored despite the fact that the systems are highly nonlinear. The unspoken assumption is that the overall system environment is a smooth one and equalizing mechanisms prevail that allow the impact of smaller scales to be neglected.

Example: What is the appropriate level of regional aggregation?

We want to illustrate the scaling problem for the choice of macro regions in a model. For most IAMs, the choice of aggregation is a mixture of geographical proximity (e.g. continent based) and similarity in energy and/or economic conditions (e.g. portfolio of resources or economic power). However, as Gruebler (2004) points out, taking the world average for analyzing the development of final energy is misleading, since drivers are substantially different for developing countries (where final energy demand is governed by growth in population) and OECD countries (where it is decoupled from population growth).

The conclusion by Gruebler (2004) can be generalized by asking the following question: what is a useful aggregation of nations to a macro region given the purpose of an IAM to study climate change impacts and the possibilities of mitigating greenhouse gas emissions? For most IAMs, Kaya's decomposition of carbon dioxide emissions into four main sources (Kaya, 1990) is a central output variable for exploring drivers of greenhouse gas emissions. An aggregation would then be useful, if important processes are not neglected. This implies that a regional choice is adequate if the regional

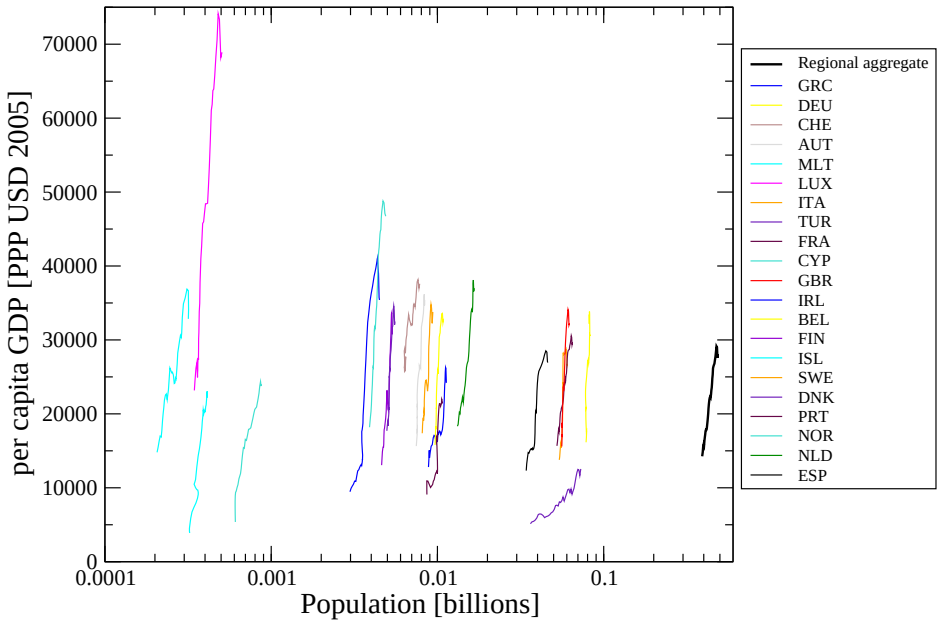

Figure 1: Western Europe as defined in GEA (2012): GDP per capita (1970-2010) is plotted against population growth. Nations are given by ISO 3166-1 alpha-3. Comparing value levels and patterns, the region is relatively homogeneous. Thus, the regional aggregate (bold-faced) is a good approximation. Source of data: ENERDATA.

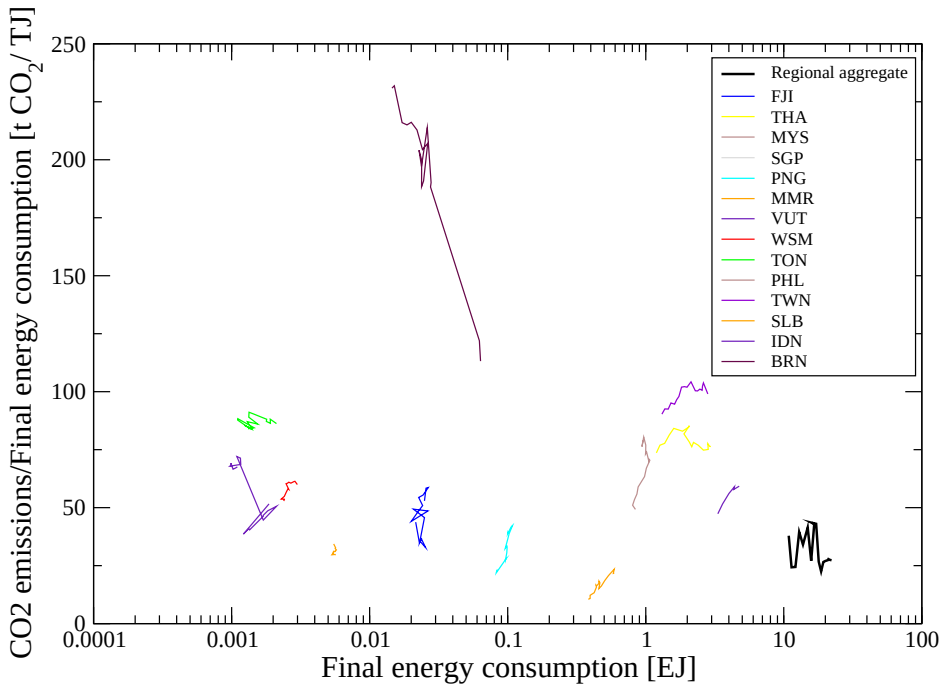

Figure 2: Pacific Asia as defined in GEA (2012): carbon intensities of final energy (1990-2010) are plotted against final energy. Nations are given by ISO 3166-1 alpha-3. The huge spread of values and patterns across nations is evident. Thus, the regional aggregate has relatively low explanatory power. Source of data: ENERDATA.

aggregate is a good approximation for nations that contribute most to a macro region. This refers to level values as well as dynamic patterns. For illustration, we choose two macro regions as defined in the recent Global Energy Assessment (GEA, 2012). We select Western Europe and Pacific Asia. Let us discuss exemplarily two Kaya-factors: GDP per capita for the period 1970-2010 as a function of growth in population, and carbon intensities of final energy (1990-2010) plotted against final energy development in that period. By doing so, we test the dependence of the Kaya-factor on its denominator. For GDP per capita, Fig. (1) shows an example of a quite homogeneous region; only Turkey stands out among the main contributors. Looking at the carbon intensities of final energy, Fig. (2) shows on the contrary that nations belonging to one macro region can be very heterogeneous - the variety of dynamic patterns and values across nations is huge. The

more heterogeneous regions are with respect to an output variable, the more likely it is that important processes are omitted. In other words, the more open a system, the lower the explanatory power of a model. Note, that an appropriate regional aggregation might also shift in time, in particular, when one of the members dominate a region.

2.3. Future behavior is fundamentally unknown

Another fundamental issue of IAMs of global climate change is the following: we are simply not able to anticipate how the future will unfold. If we knew how the future would develop, we would adapt our expectations and subsequently our activities. This would in turn change what was beforehand considered to be the future. Therefore, empirical data of the system will only be available in the future and no experiment can be set up to generate them in advance. If the reference value is not available today, the confrontation of IAM results with empirical data from the real system is only possible in retrospect. This however is a dilemma (for all time evolving systems), as already today we need to assess how much model results can be trusted. In addition, integrated assessment modellers have to deal with potential changes in qualitative system dynamics, e.g. by the appearance of new sub-systems or processes that are currently not existing or are completely unknown.

Fig. (3) underlines the difficulty of projecting what is fundamentally unknown (see e.g. Smil (2000) among others for a retrospect evaluation of U.S.A. energy demand forecasts). The figure shows how various projections of world electricity demand are complying with actual data and, furthermore, how projections have been adjusted with time. Historical data for electricity demand are coming from ENERDATA and denoted by crosses. Projections are taken from the World Energy Outlook (WEO), Exxon Mobil (2012), and Institute of Electrical Engineers Japan (2011). Baseline projections of the IAMs MESSAGE and IMAGE as used in GEA (2012) are also included. There are two main observations: first, there was a systematic overestimation of world electricity demand by WEO projections (1994-1996) and an upward correction for 2020- and 2030-projections. Second, the spread across projections increases with each decade looking further ahead. For most recent projections, the spread sums up to roughly one fifth of the total demand by 2040 - which is only three decades from now.

Given the fact that the time perspective of many IAMs is not a few decades but as much as 100 years into the future, it is obviously useful to increase efforts on model evaluation and to think about the purpose of IAMs.

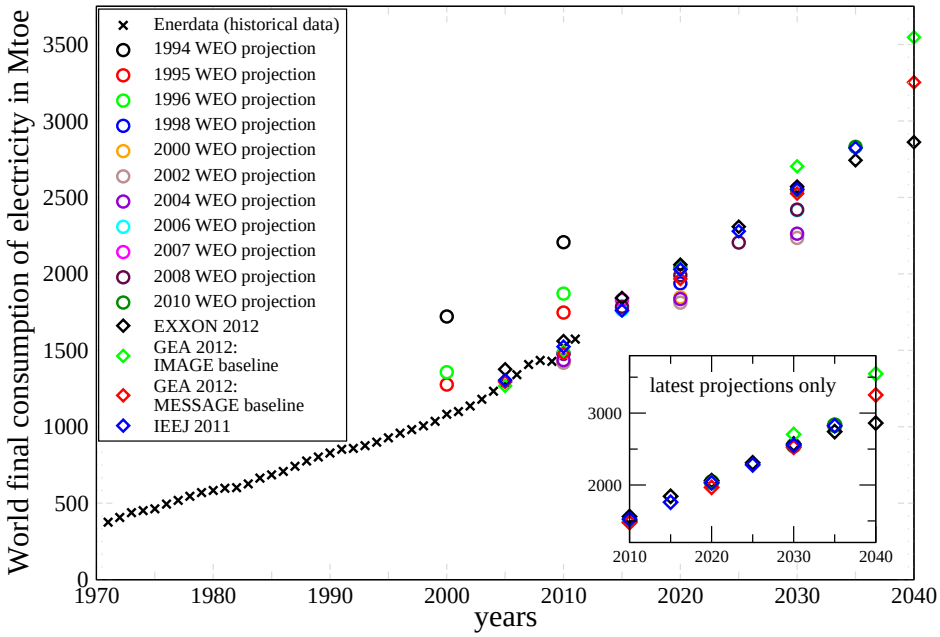

Figure 3: Projections of world electricity demand by WEO (1994-2010), Exxon Mobil (2012), Institute of Electrical Engineers Japan (2011), and GEA (2012). Historical data are from ENERDATA. Projections show an over-estimation of electricity demand in the 90s. The spread across projections increases the further the models look into the future.

As the retrospect analysis revealed, anticipation of system shifts has failed (Smil, 2000) - forecasts simply did not work. As a consequence, integrated assessment modelling in climate change concentrates on scenario exploration (if-then-analysis) where relative changes in view of the integrated system are studied instead of numbers. In other words, the research focus is on the explanatory power of a model instead of aiming for quantitative forecasts.

3. Aspects of model evaluation

3.1. Evaluation as a continuous effort

If confidence in knowledge about the complex system is only revealed in a semi-formal process, the evaluation of IAMs becomes a matter of step-by-step confidence building. This could also be understood as paying tribute to the "prolonged nature of model validation" (Barlas, 1996). Such an approach has the advantage that views of different groups can - indeed have to - be

integrated. These include, apart from opinions in the scientific community, the perspectives of stakeholders and external experts as secondary users of model results. This inclusion means that stakeholder needs can be better reflected on the one hand and the trust of third parties in the model can be increased on the other. It underlines that "usefulness" is an integral part of evaluating large and complex models - an opinion that is shared among different modelling communities, e.g. Landry and Oral (1993) for Operations Research, Sargent (2003); Jakeman et al. (2006) for the broad class of simulation models, Oreskes et al. (1994) for earth system models, and Cash et al. (2003) for knowledge systems in general. Important dimensions of usefulness of an IAM are the usefulness of the purpose itself, the manageability of the model, as well as the usability of given answers.

A second consequence of a scientific discourse characterized by loops and feedback is that evaluation routines also need to address the process of model development, see Jakeman et al. (2006). For this purpose, coding etiquette and guidelines for good modelling practice are constructive. Such formal parts of model validation, including the exclusion of coding errors and mistakes in implementing the conceptual model, are strongly emphasized in Fisher (2007) and Oberkampf and Roy (2010).

Finally, the evolutionary character of the evaluation process calls for applying a variety of tests to learn about the model questioning its behavior. This pluralism of evaluation methods seems the only way to balance the fundamental problem that experimental data of the real system do not exist and that just confirming the historical data by turning the IAM clock back does not help either. Developing and adopting community standards and, possibly, performance indicators is valuable as a means to increase transparency and consensus among all stakeholders in this context.

3.2. Connection to a model's purpose

Linking model evaluation and usefulness, a model's fitness for purpose becomes an important element (see Barlas, 1996; Risbey et al., 1996; Beck et al., 1997; Jakeman et al., 2006; Weyant, 2009, and citations therein). This requires that the purpose of the model is clearly specified, including its intended use and group of users. Burton (2003) classify model objectives by the type of research questions that are asked. 'What is'-questions are answered by analytical, descriptive models. In consequence, the degree of realism becomes a focus in evaluation exercises of these models. A second

class tackles 'What might be'-questions. Here, the objective is not to reproduce a real-world-situation but to explore a range of possible, i.e. imaginable, realizations of the dynamic system. Hence, evaluation becomes a matter of judging how reasonable different model answers are. A normative element is added if 'What-should-be'-questions are investigated. As preference choices are involved, the inclusion of all groups of users (e.g. policy-makers) adds to the credibility of evaluation exercises. Admittedly, IAMs of global climate change share characteristics of all three classes, suggesting that an evaluation framework should be build upon various tests and exercises.

The main evaluation question that needs to be asked is: can we confidently apply the model to deliver a well-grounded answer to the group of users? For an answer to be well-grounded, three aspects are necessary but not sufficient. One is that the omission of important processes and phenomena with respect to the intended research questions is kept to a minimum and the rational for prioritizing is defensible. On the other hand, "complexity is a threat to construct validity" (Burton, 2003). Therefore, Saysel and Barlas (2006) suggest a framework to iteratively simplify models as part of the evaluation procedure. The second aspect concerns the extent to which this wish-list of important processes and phenomena is actually implemented in the model.[2] Again, the rationale for choosing the level of detail needs to be scrutinized. These issues are best addressed in a series of workshops involving different groups, including those who are intended to use a model's output. The aim of these workshops is to shape a common understanding on what is captured in the model - or what not. Whereas the focus of the first two purpose dimensions is on a model's scientific legitimacy, the third aspect underlines prominently the intended use of the scientific knowledge. Cash et al. (2003) emphasize the role of "credibility" (from a science perspective), "salience" (fitting the need of users), and "legitimacy" (integrating different views).

3.3. Testing the model's structure and behavior

As the integrated system perspective distinguishes IAMs from stylized theoretical or empirical models, it is not only of interest what comes out of the model but also why; in other words, IAMs of global climate change are

[2]Risbey et al. (1996) summarize three assumptions that IAMs are based on: all relevant phenomena can appropriately be modelled, a modelling paradigm such as system dynamics is suitable, and results are policy-relevant.

in the group of cause-descriptive (white-box) models. This cause-descriptive claim shifts the focus from only output or behavior evaluation (are the quantities right and accurate?) to testing the plausibility of the model's internal structure (is the model's story right?). However, confidence in the structure does not guarantee correct results. Hence, structure evaluation tests have to be followed and complemented by behavior tests. This view is shared by other communities, e.g. Barlas (1996) suggests having "technical tests gathered around the theory base of system dynamics"; see furthermore Konikow and Bredehoeft (1992) as well as Huntington et al. (1982) and Schneider (1997) stressing the value of "insights". In addition, the fundamental problems of IAMs discussed in the previous section leave no choice but to scrutinize the plausibility of the model behavior and their structure for evaluating them. A change in the focus of evaluation towards plausibility and insights has strong implications. First, evaluation routines can only assess the explanatory power of an IAM, not its forecasting power. The focus is on tests that evaluate model structures. Examples are model walkthroughs, behavior sensitivity analysis, or an analysis of the relationship between variables (see also Barlas (1996) for a comprehensive list). Second, it is scales, patterns, and processes that matter, not singular events or a statistical matching with time series. Third, as research restricts itself to an if-then-analysis, applying model results without relating them to a reference case is impossible.

Given the complexity of the system, it is useful to build an evaluation hierarchy comprising the complete system, sub-systems, as well as benchmark cases. This is similar to what Risbey et al. (1996) call "discipline based", "process based", and "end-purpose based" assessments. Various diagnostic tests can be designed to evaluate the structure of the model, its parts, and their inter-linkages (see Section 4 for examples). Putting these tests in the context of other IAMs' behavior enriches the exercise. Useful insights can furthermore be gained by confronting the causal chain of the model output with historical data or stylized behavior patterns generated from robust observations. Instead of judging the model on the basis of compliance or non-compliance, the idea is to explain and defend in a scientific discourse whether the differences are reasonable or not. Such tests gain momentum if done systematically, e.g. building on a collection of widely-acknowledged stylized behavior patterns relevant to the system an IAM aims to describe.

3.4. Evaluation does not work without documentation and communication

Transparent documentation and targeted communication are crucial towards building trust in the model and its results. The question is therefore: what kind of information is essential and how can it best be passed on?

The package of essential information should include at least the description of the model (specification of its purpose, the conceptual/mathematical model, assumptions and sources of data) and a documentation for developers about implementing and operating the model as well as a documentation for users of model results (analyst and stakeholders). Blueprints for documenting a model are an advantage as these make it easier for third parties to filter for relevant information; supportive are also examples of how model results can be interpreted or misinterpreted. Furthermore, it needs to be documented why it is legitimate to use the model results. This can be demonstrated by documenting the evaluation process (for each research question the model addresses) and publishing conclusions from performance tests whose credibility can be increased if they adhere to community-wide standards. For example, it could become common practice to document reference scenarios (business as usual, policy baseline, 450 ppm and 550 ppm climate stabilization targets) in a shared database. A useful format to document the model and the results of the evaluation process are, e.g., evaluation tables as suggested by Sargent (2003, 2010) or phenomena identification and ranking tables (Oberkampf and Roy, 2010), see also Section 4 for an illustration.

R.G. Sargent - a pioneer of model development and validation - also suggests the use of a confidence measure in the evaluation tables. While he "does not believe in the use of scoring models" (Sargent, 2010), he proposes ordinal scales to record the confidence in results. Renn and Levine (1991) see trust as being established if the message is being transmitted "accurate", "objective", and "complete" - evaluation in the end also means being accountable for the public (Jasanoff, 2010). The author describes this as a "three-body-problem" involving "scientists, scientific knowledge, and committees translating science into policy relevant forms". Approaches on how to organize the involvement of different groups are discussed in the literature, see e.g. Voinov and Bousquet (2010); Krueger et al. (2012) for reviews. They range from indirect forms such as the publication of quality criteria (Beck et al., 1997) or suggestions on how to improve communication of uncertainties (Budescu et al., 2009) to direct forms such as the use of iterative dialogues (Parker et al., 2002) and participatory approaches or the inclusion of expert opinions (van der Sluijs, 2002; van der Sluijs et al., 2005; White et al., 2010).

In addition to purely informing third parties and considering their views and opinions, communication in the context of evaluation also serves the purpose of mitigating misinterpretations or errors by users of model results (Beck et al., 1997; Brewer and Ley, 2013).

Recently, the stronger involvement of stakeholders in the IAM development process is a hot topic in the larger IAM community. A comprehensive review of the literature is Krueger et al. (2012). They argue that even "the technical process of modelling itself can be subject to stakeholder scrutiny and input of stakeholder expertise, and can lead to the generation of new knowledge." At the same time, they acknowledge that "areas such as climate modelling are so removed from non-specialist experience that they do not lend themselves to inputs from non-specialist experts."

4. A proposal for an evaluation framework

We summarize the discussion above by proposing an evaluation framework for IAMs of global climate change. A main point is that continued efforts are necessary in order to test a model's performance. No model can be once and for all certified as 'valid to use' since model development is continuing and step-by-step more knowledge about the system and its behavior is revealed. While focusing on the explanatory power, the model needs to be assessed in parts (i.e. its purpose, its input and output, modules and components) and as a whole. To cope with the fundamental problems in validating IAMs (repeating: the openness of their systems and the lack of real system data), the semi-formal evaluation process should build on a pluralism of methods comprising as many different tests as possible (comparison to historical data and dynamics, inter-model comparison, confrontation with expert opinions etc.) and involving different parties (incl. non-expert end-users). IAMs of global climate change passing all steps in such an evaluation hierarchy are more trustworthy addressing a range of research questions in compliance with the model's purpose than those not. The formal result is a statement of confidence in a model's usability. Documentation (incl. limitations and gaps) and communication are integral parts of evaluating a model.

We suggest that the evaluation hierarchy for IAMs should comprise of loops that differ in intensity of transversing them. The full cycle includes setting up an evaluation framework, scrutiny of the conceptual model, code verification and model documentation, model performance tests, uncertainty and sensitivity analysis, documentation of the evaluation process, as well as

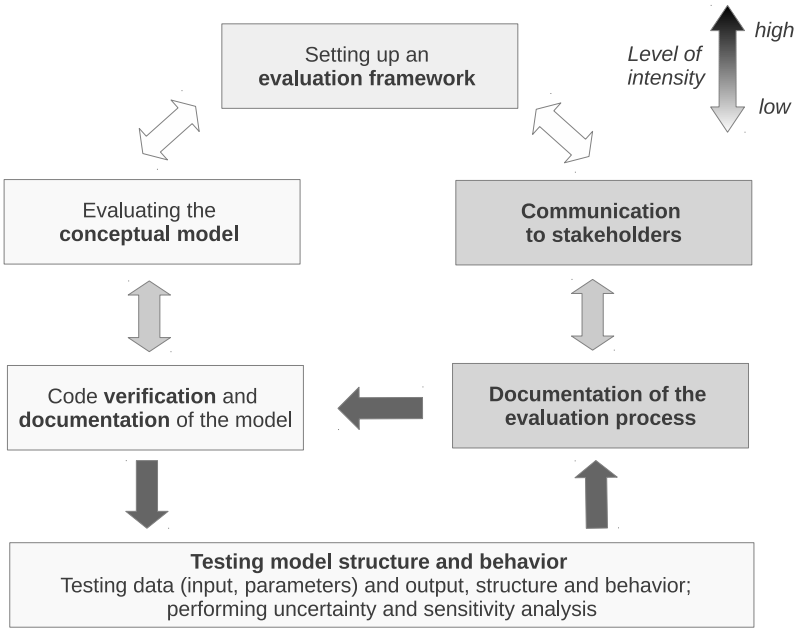

Figure 4: The hierarchy of evaluation exercises comprises of loops that differ in intensity of transversing them (compare color code of arrows).

communication to stakeholders (compare Fig. 4). For existing models, the full cycle may only be performed at larger intervals, whereas the loop from code verification to evaluation, uncertainty and sensitivity analysis as well as its documentation should be repeated more often depending on the pace of model development. It should be pointed out that the proposed hierarchy has some steps in common with the ten steps of building a model in Jakeman et al. (2006). The hierarchy in our paper, however, focuses on a specific class of models, those addressing global climate change. Furthermore, these models need not to be built up, but are already established. Therefore, some of the 10 steps are less relevant for our purpose. Additionally, the last step in Jakeman et al. (2006) "Model evaluation or testing" needs to be adopted and extended for the class of models we are looking at. In the following, the content of the single steps in our evaluation hierarchy are discussed.

Setting up an evaluation framework

A formal agreement should be reached within a team and among the modelling teams on how to evaluate models and which evaluation exercises to undertake at what level of frequency. It is important to establish a formal scoring system ranking the different evaluation steps and exercises in view of the model's purpose. The state of art in the community can provide useful guidance on how to design an evaluation hierarchy and a scoring system for an individual model. The establishment of a scientific working group for such purposes in the IAMC[3] is a promising starting point for developing community guidance. Useful approaches can also be found in documents developed within other communities, see e.g. model quality checklists, Risbey et al. (2001), and good practice guidelines, Ravetz (1997). Regarding concepts and methods in scientific computing in general see Oberkampf and Roy (2010) among others.

Establishing an evaluation framework requests substantial efforts and resources. First lessons learned from starting this process with the model REMIND[4] are:

- How to finance? Third-party funding can provide resources. It is of mutual interest to stress evaluation as an indispensable part in project pro-

[3]IAM Consortium, <www.globalchange.umd.edu/iamc>.

[4]See <www.pik-potsdam.de/research/sustainable-solutions/models/remind>.

posals. For example, the AMPERE inter-model comparison project[5] includes a work package dedicated to advancing methods of model diagnostics and validation. The project is furthermore well linked to the community of IAMs.

- How to get started? Using institutional retreats to brainstorm and discuss about model evaluation in a group larger than the modelling team. This allows to include scientific expertise from other modellers and to link model evaluation to related topics, e.g. model realism, interplay between research and policy advice, and up-coming research priorities.

- How to maintain activities? Establishing joint and regular (e.g. quarterly) 'model service days'. By dedicating a whole day to tasks such as code cleaning, model documentation, and/or model evaluation, etc., joint responsibilities and group ownership can be re-enforced and collective learning is fostered.

Another idea helpful for getting started to set up an evaluation framework is to document the tacit knowledge of model developers and model applicants. The aim is to record knowledge that is often lost, e.g. because people move on to other research topics or institutions. In the REMIND team, interviews have been conducted asking each team member about her or his insights on working with the model. Results were fed into a wiki-based model documentation system. The questions ranged from running the model to checking and interpreting model results. Another purpose of this exercise was to make explicit and document parameters and bounds that are based on rules of thumb of a modeller. Interviewees could also add to a list of tests that could/should be undertaken if time allowed. They were also asked to point towards gaps and shortcomings of the model from their perspective.

Evaluating the conceptual model

The evaluation of the conceptual model ideally starts with taking a step back from actual model implementations. What is needed is a reflection about a model's purpose and related overall research questions: Is the model purpose specified and documented? Is there a joint understanding in the

[5]AMPERE, <www.ampere-project.eu>.

Model objective: Qualitative assessment of mitigation options for global climate change

System process	**Sub-processes**	**Representation in model X**
Social & cultural change	Individual behavior	not considered, aggregation to macro regions
	Social choice	representative agents and rationality paradigm, assuming an exogenous global discount rate
Institutional change	Rules, regulations, law	assuming perfect markets and free trade
	Policies, government	intertemporal welfare optimization, explicit technology and climate policies
Sudden events	Luck	not accounted for, normal system environment
	Catastrophes	
Geographic change	Resources, reserves	exogenous, coverage of fossils (coal, gas, oil), solar and wind with regional potentials
	Land-use patterns	using an emulator (land-use model Y)
Climate change	Global patterns	endogenous global mean temperature
	Regional patterns	e.g. sea level rise not covered
Economic development	Economic growth	tuning to exogenous GDP scenarios, calibration of labor efficiency, endogenous investments
	Demographic change	tuning to exogenous population scenarios
	Structural change	specification of urban/rural population, resolving 5 end-use sectors (shares are endogenous)
Energy transition	Change in quantities	tuning to exogenous energy demand scenarios
	Change in qualities	endogenous, 50 transformation technologies incl.
	Change in structure	endogenous, calibration of base year costs
Technological change	Creation of knowledge	single generic, speculative back-stop technology
	Diffusion of knowledge	learning by doing, no explicit knowledge stock, technology diffusion driven by relative costs

Table 1: Example of a PIRT documenting relevant system processes and their coverage by an IAM of global climate change. See also extensions of the sketch for documenting the evaluation process (Tab. 2).

modelling team? Has the research focus broadened or shifted? What follows is an assessment on how well the conceptual model is in line with the model's purpose, allowing to address the intended research questions. A systematic and formalized review is crucial. Useful methods for doing this are brainstorming exercises and other methods supporting the identification of gaps and shortcomings of the conceptual model, see e.g. Oberkampf and Roy (2010, chapter 14), van der Sluijs et al. (2005).

A possibility for a systematic assessment is the use of phenomena identification and ranking tables (PIRT), see Oberkampf and Roy (2010). The basic principles can be adopted to suit IAMs. Tab. 1 sketches this idea: Starting with a model's objective and overarching research questions, a modelling team (possibly extended by experts and stakeholders) pins down relevant system processes. This includes the specification of sub-processes, time scales, and measurable quantities. A ranking of processes in line with their relevance for the model's purpose is important. What follows is an assessment on how these processes are represented in the model, e.g. are they exogenous assumptions, endogenous model results, or processes not covered, etc.? This allows to identify shortcomings in the conceptual as well as procedural model and its structure. The table can furthermore be linked to results obtained from other steps in evaluating the model (see Subsection 'Testing model structure and behavior' and the example provided therein).

Code verification and documentation of model

The model code needs to be verified in comparison to the conceptual model; the numerical solution algorithm needs to be reviewed and tested. Helpful tools for these purposes are checklists, e.g. Risbey et al. (2001), and formal coding etiquettes (handling of naming and unit conventions and commenting, application of four-eyes principle, version control system, online discussion platforms, etc.).

The model needs to be transparently documented for an internal and external audience (for the latter see also following sections). The documentation comprises commenting in the code, model description including specification of its purpose, data, and assumptions, instructions for operating the model, and archiving data. The IAM community can provide support to modelling teams by developing coding etiquettes, good-practice guides, as well as blueprints for model description.

It is a characteristic of IAMs that the model is never finalized; constant extensions and refinements are implemented to address real world processes.

System process	Stylized behavior pattern	Model evaluation & conclusion
Social & cultural change	Empirical findings on herding behavior	extend model for social choice features, e.g. based on contagion models
Institutional change	Role of informal economy, e.g. as share of income	explore market failures, asymmetric knowledge, or myopic behavior
Sudden events	Frequency of oil shocks	develop shock experiments, design of hind casting experiment
Geographic change	Empirical findings on speed of land conversion	test revealed: speed was overestimated, revision of module necessary
Climate change	Observation of global warming	test alternative climate models
Economic development	Speed of urbanization	violated by choice of spatial aggregation, low confidence in regional results
Energy transition	Distribution patterns of energy efficiency	coherence with alternative long-term projections found
Technological change	Speed of technology diffusion	spread across regions unrealistic for offshore wind, outlier in community

Table 2: Framework to evaluate IAMs of global climate change (expansion of Tab. 1).

A well documented model assists the learning process of modelling by also showing how a model's story develops across revisions.

Testing model structure and behavior

The evaluation exercises should target input, output, structure, and behavior of the model as a whole and its components. The menu of useful tests includes information flow analysis, inter-model comparisons, diagnostic tests, decomposition-analysis of output aggregates, blind-testing with external experts (Turing tests), checks against historical patterns and trends, as well as hind casting exercises. Complementing the tests of the internal model structure, model results need to be confronted with other sources of experimental data and/or knowledge, preferably with data not used for calibrating the model. A comprehensive overview on tests of formal model evaluation is e.g. given in Fig. 1 in Barlas (1996) who starts the evaluation procedure with testing direct structure and structure-oriented behavior tests followed by behavior pattern tests. In the following, we briefly discuss selected tests that seem particularly useful for IAMs.

Model inter-comparison exercises are an important tool to assess model structure. The Energy Modeling Forum[6] initiated a long tradition of such exercises in the IAM community (Sweeney, 1983). In recent years, community projects gained popularity fostering in turn the establishment of necessary infrastructure to carry out such extensive tasks. While most of the studies

[6]EMF, <http://emf.stanford.edu/>.

are science and policy driven, they also serve as a mean to compare and understand model differences. Model inter-comparison exercises specifically targeting diagnostics and evaluation are, however, still in its infancy. In these diagnostic tests, a set of carbon dioxide price trajectories is defined and used as prescriptions to the models. This practice can be extended by shock or extreme condition experiments to explore model boundaries and scan the solution space. Results are particular insightful if model input is harmonized between participating models. Again, standard diagnostic experiments and indicators should be agreed on in the community.

Given the lack of future data to compare model results with, patterns and trends observed in history (stylized behavior patterns) can be used to reveal implausible behavior in future projections. Two examples are discussed in the section below, see also Wilson et al. (2012) for a further example. The value of such confrontations with stylized behavior patterns increases if done systematically. For this purpose, the PIRT can be extended (see Tab. 2): System processes and sub-processes are complemented by empirical knowledge about the system. The link to the representation of processes in a model is thereby established. This offers the possibility to design tests for evaluating the model. The aim is to assess on the one hand, whether the transition from history to future projections is smooth. In this way, problems in calibrating a model's base year and in tuning a model to exogenous scenarios can be identified. On the other hand, a quantitative and qualitative judgment about realistic or unrealistic system behavior of the modelled future is possible. The results and conclusions from the comparison with stylized behavior patterns can also be included in Tab. 2. Finally, coming up with a list of stylized behavior patterns and agreeing on it in the community would increase comparability across the models.

Hind casting is another mean to learn about the plausibility of a model. In these tests, the clock of a model is turned back (say to 1945) and it is tested whether the model can reproduce developments in history (say global energy trends in the last decades of the 20th century). However, being in line with historical developments does not provide confidence that the model can also describe the future well. Furthermore, IAMs are designed to carry out if-then-analysis. The analysis' set-up includes a baseline scenario (i.e. the continuation of current trends) and alternative scenarios (i.e. exploring trajectories of change). Note that there is no assignment of likelihoods to the scenarios, in other words: the models are not designed for forecasting.

Uncertainty analysis and sensitivity analysis are important tools to access

the robustness of a model in view of its input data, parameters, resolution level, and model structure. While uncertainty analysis tries to quantify the uncertainties in these elements and their propagation by the model, sensitivity analysis aims to attribute the uncertainty in the output to the uncertainties in the input (Saltelli et al., 2008). Sensitivity analysis might help to encounter unexpected relationships between inputs and outputs, lead to model simplification, or reveal simple relationships in the story told by the model and thus supports communication of model results.

Advanced forms of sensitivity analysis are variance-based approaches and mainly due to Sobol'. Here, one decomposes the output variance with respect to the input variances (Sobol', 1993). It is often combined with Monte Carlo sampling (Sobol', 2001) of the multi-dimensional input space. Note however, that the study in the annex of the Stern report, Stern (2007), has been criticized by Saltelli and D'Hombres (2010) for its limited parameter space. This is typically the biggest challenge in performing a sensitivity analysis: large number of parameters and long run-times severely limit the exploration of the entire input space. To a very reduced scale, diagnostic runs can operate as a simplified sensitivity analysis. It should be mentioned that sensitivity analysis as well as uncertainty analysis of the model structure are seldom performed. Only few general concepts are available to carry out such a task. To some extent the different set-ups of the IAMs already give an indication of uncertainties in the model structure once key inputs and parameters are harmonized. Results of the analysis of sensitivities and uncertainties can also be documented by adding a further column to the PIRT, compare Tab. 2.

Example: Evaluation with stylized behavior patterns

We want to illustrate how to evaluate integrated assessment models of global climate change using stylized behavior patterns in combination with the framework sketched in Tab. 1 and Tab. 2. We first describe the stylized behavior patterns as it is observed in historical data. Next, we discuss how system processes (linked to the stylized behavior pattern) are represented in the models (input to Tab. 1). This is followed by a confrontation of the historic plot with the plot obtained from model results. Finally, we draw a conclusion from the comparison (input to Tab. 2). We choose two stylized behavior patterns, showing examples for the system processes "Economic development" and "Energy transition" (refer to Tab. 1). As in Section 2, we use results from GEA (2012).

The idea of using stylized behavior patterns has been developed by Kaldor

(1961). He refers to "stylized facts" in the context of building useful economic growth models. Kaldor suggested that these models should be able to reproduce six stylized facts, known as Kaldor's facts. He defined them as "broad tendencies, ignoring individual detail" and "characteristic features of the economic process as recorded by experience". On the occasion of their 50th anniversary, Kaldor's facts have been revisited and updated by Jones and Romer (2010) (New Kaldor facts). Recently, the concept has been transferred to complex social systems in general by stating that stylized facts are "mathematical patterns suggesting some deep order within some of our most chaotic social systems" (Buchanan, 2012).

The first stylized behavior pattern we discuss is the New Kaldor fact No. 3 (Jones and Romer, 2010). It states that "The variation in the rate of growth of per capita GDP increases with the distance from the technology frontier." This can be inferred from the smaller inlay in Fig. 5, where average growth rates (observed across different countries within four decades) are plotted against the logarithm of GDP per capita in the base year 1960 (refer to the growth triangle and note the normalization to the technology frontier). How are the sub-processes "economic growth" and "demographic change" represented in IAMs of global climate change (refer to Tab.1)? In most IAMs, both are exogenous input. This is also the case for the models used in GEA (2012), which are tuned to fit exogenous GDP and population scenarios described as "median economic development paths ... consistent with global aspirations toward a sustainable future" (GEA, 2012, chapter 17, p. 1221). We show in the main plot of Fig. 5 how New Kaldor Fact No. 3 unfolds in the period 2010-2050 driven by the scenario assumptions. We observe that the fact is not reproduced as countries show no spread for large distances from the technology frontier, i.e. at small GDP per capita. As we do not know how the future unfolds we can not say that this strong convergence bias is right or wrong. However, in view of historic developments it seems rather optimistic (e.g. COD moving from a de-growth path to an annual average growth path of 7 %). We conclude that it would be necessary to develop alternative GDP scenarios testing the robustness of main findings and the impact on economic development at the level of regional aggregation.

The second stylized behavior pattern we want to briefly discuss is one of the most relevant facts of the energy system: economic and energy growth show an "overall positive correlation, that is, however, variable over time" (GEA, 2012, Chapter 1, p. 115). Fig. 6 illustrates this stylized behavior pattern at the level of primary energy per capita plotted against GDP per

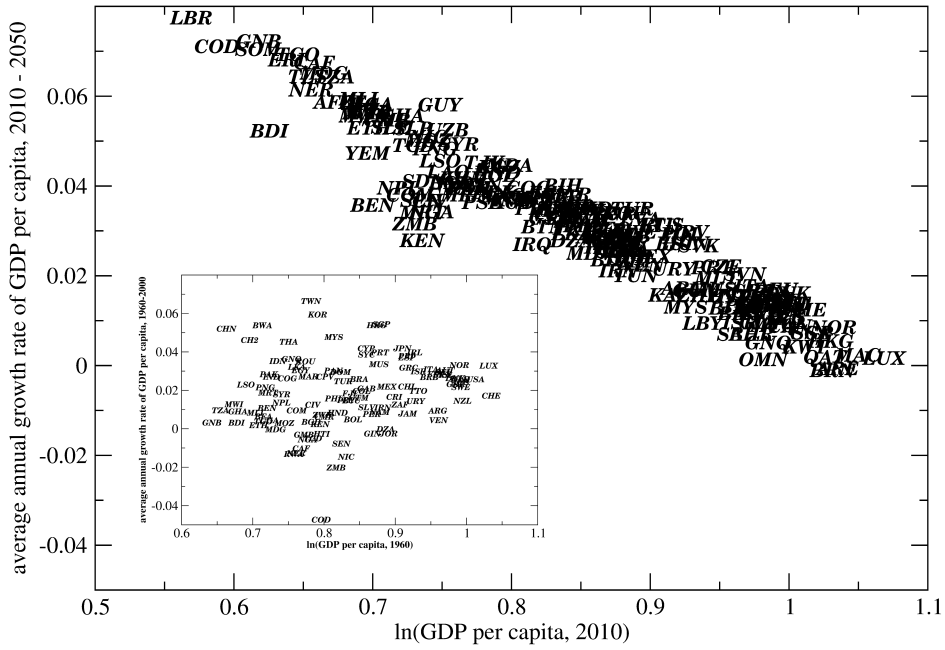

Figure 5: New Kaldor fact No. 3 states an increased variation in growth with growing distance from the technology frontier (normalized to USA). Annual average growth rates of GDP per capita for four decades are plotted against the logarithm of GDP per capita in the base year. The small inlay shows the stylized historic growth triangle. The main plot shows projected model results. Source of historical data: Heston et al. (2006). Source of model data: GEA (2012).

capita. Historical data are shown at the country level for USA, Japan, and India (solid lines) and for the aggregation to GEA regions North America - NAM, Pacific OECD - PAO, and South Asia - SAS (dotted lines). As the differences between country level data and regional aggregation are similar in historic data, we conclude that the choice of aggregation is useful in view of the variables. Contrary to the example discussed above, primary energy is an endogenous result of models used in GEA (2012). Confronting historic patterns with projected model results, we observe that the overall positive correlation is pertained for the regions SAS, NAM, and the world aggregate. However, for the case of PAO primary energy consumption per capita is rather independent from GDP per capita for a longer period (2020-2050), whereas there is a strong increase for later decades. We conclude that it is necessary to have a closer look into this region. The purpose of the exercises would be to reveal the plausibility of the observed deviation from the stylized behavior pattern and the later sudden change in behavior, e.g. by supplementing the evaluation with the use of other related stylized behavior patterns.

Documentation of the evaluation process

Essential is a summary of the blocks that have been carved to build trust in the model. The summary should include a formal statement on what can be delivered by the IAM tested and at which level of detail (applicability domain); gaps are explicitly named. An example for such a framework is the NUSAP-system (Funtowicz and Ravetz, 1990; van der Sluijs et al., 2005). We have furthermore discussed the PIRT which can also be used for this purpose. The structure makes it easier to compare documentation across models. Also, it is easy to grasp by non-specialists and the folder-like format of the table is accessible to web-based content-management systems.

It is useful to include some kind of a community-wide accepted performance indicator and/or 'objective' measure. This can lead to a strength-weakness analysis. The results can be summarized with visualization techniques, e.g. spider-diagrams, kite-diagrams, tables, etc. The ranking (scoring) of different methods is up the prior agreement of a modelling team or the community.

In this course, models can be categorized in accordance to the outcome in evaluation exercises. Models taking higher ranks can claim a higher confidence level w.r.t. model evaluation and model transparency. The performance of models could become a valuable consideration in the peer-review

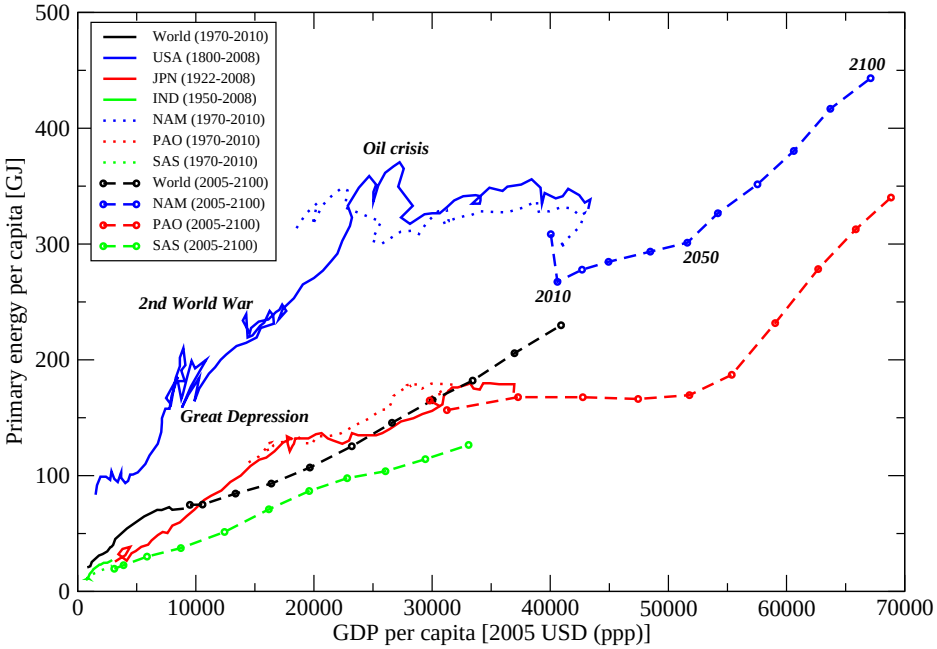

Figure 6: Example for stylized behavior pattern related to energy transition: There is an overall positive correlation between per capita primary energy consumption and per capita GDP. Source of historical data: GEA2012 (for countries) and ENERDATA (for regional aggregation). Source of model data: GEA (2012).

process. For example, it could become a standard in publications to also provide diagnostic and evaluation indicators that put a single model in relation to other IAMs. The modelling team (or the community) should agree on thresholds for models performing poor and models that are below the threshold would be rejected. For example, models with insufficient documentation should be rejected on this basis. However, in defining the threshold one should have in mind the statement of Beven "that it might be dangerous to exclude models that are consistent with observational data but lead to unexpected conclusions in prediction" (Beven, 2002).

Communication with stakeholders

A passive element is the provision of an essential information package supplying sources of model documentation and user guidelines, a summary of the evaluation procedure, as well as a statement of confidence in the model. A good practice example to increase transparency is the open source policy of the models GCAM and DICE, see websites GCAM (2012) and Nordhaus (2012). Another extremely valuable development is the standardization of model outputs. Driven by model inter-comparison exercises, standard output templates for some hundred variables have been developed. An increasing number of final model output is collected in a community database at IIASA which is accessible to the public. Useful are also interactive websites for hand-on exploration of model results and scenarios. An example is the model simulator of WITCH[7].

Dissemination to stakeholders and the public can be actively supported by face-to-face communication, e.g. discussing examples of how output variables might be interpreted or misinterpreted. Common practice are the involvement of non-specialist experts and model end-users in project steering committees (via participation in project workshops and the provision of suggestions to the modelling teams) and the organization of stakeholder conferences to bring project messages across. An example for the inclusion of experts from other disciplines to review model input data is Bosetti et al. (2012). In this paper, expert elicitation is used to refine process understanding and data used for representing solar technologies in IAMs. An example to include domain-experts is the RoSE project[8]. For a review of concepts

[7]See <www.witchmodel.org/simulator/>.

[8]See <www.rose-project.org/consortium>.

on expert elicitation see Krueger et al. (2012). Inspiration for participatory approaches can be found in Voinov and Bousquet (2010); White et al. (2010). Kloprogge et al. (2011) is an example for assessing the value-ladeness of model parameters, assumptions, and the influence of model-end users on the modelling process.

5. Conclusions

The intention of this paper is to feed the discussion on how Integrated Assessment Models of global climate change can be evaluated. The main points are, firstly, that evaluation of IAMs is best understood as a step-by-step demonstration of a model's usefulness, in particular the plausibility of its behavior. This gradual process of performing a variety of tests needs to be accompanied by an open discourse from which a robust consensus can crystallize. A transparent documentation and communication is integral to the evaluation process.

The second point concerns the value of developing community-wide acknowledged standards for performance testing. This would in particular increase inter-subjectivity (Krueger et al., 2012) and enhance transparency. A predestined forum for developing, discussing, and adopting a joint evaluation framework is the Integrated Assessment Modeling Consortium, IAMC. A promising step is the plan of establishing a scientific working group for this purpose. However, such an effort is a matter of years and needs continuous funding, which is a lesson learned from experience in the climate modelling community (see e.g. Covey et al., 2003).

Supported by reviewing evaluation aspects in comparison to their understanding in other communities, we have proposed an evaluation hierarchy for IAMs of global climate change. This included a discussion how the process could practically be organized. From our perspective, action-items are:

- To assign enough resources to evaluation exercises in the budget of research projects and/or to establish a community fund or program for this purpose.

- Develop guidelines for good IAM modelling practice covering the evaluation hierarchy, as described e.g. in Fig. 4. Suggestions of community-wide standard diagnostic and evaluation tests are useful.

- Discuss and test a list of relevant historical patterns that could be used to evaluate model structure and behavior and is shared in the community.
- Design a blueprint for an easy accessible chart summarizing a model's weaknesses and strengths found in the evaluation process. The aim is to set a standard for transparency and comparability across models.
- Discuss what information about model evaluation should be offered to peer-reviewers in addition.

Although community standards are of immense value, it should be kept in mind that IAM are entering a Terra Incognita for many research fields. If our knowledge about the isolated systems is already poor, it is even poorer for the integrated, complex system. Thus, the development of performance standards needs to account for the unavoidable trade-off between the indispensable trial-and-error in research and the legitimate request for providing useful answers to decision makers and the public.

Acknowledgment

The author would like to thank the anonymous referees, as well as Michael Flechsig, Jan Philipp Dietrich, Markus Bonsch, Elmar Kriegler, Michael Jakob, Eva Schmidt, and Alison Schlums for valuable suggestions in improving the manuscript. The manuscript also benefited from discussions with participants at workshops of the AMPERE and PIAMDDI project.

The research leading to these results has received funding from the European Union's Seventh Framework Program [FP7/2007-2013] under grant agreement n° 265139 (AMPERE).

References

Barlas, Y., 1996. Formal aspects of model validity and validation in system dynamics. System Dynamics Review 12, 183–210.

Barlas, Y., Carpenter, S., 1990. Philosophical roots of model validation: Two paradigms. System Dynamics Review 6, 148–166.

Beck, M.B., Ravetz, J.R., Mulkey, L.A., Barnwell, T.O., 1997. On the problem of model validation for predictive exposure assessments. Stochastic Hydrology and Hydraulics 11, 229–254.

Beven, K., 2002. Towards a coherent philosophy for modelling the environment. Proc. R. Soc. Lond. A 458, 2465–2484.

Beven, K., 2004. Reply to 'The emergence of a new kind of relativism in environmental modelling: a commentary' by Philippe Baveye. Proc. R. Soc. Lond. A 460, 2147–2151.

Bosetti, V., Catenacci, M., Fiorese, G., Verdolini, E., 2012. The future prospect of PV and CSP solar technologies: An expert elicitation survey. Energy Policy 49, 308–317.

Brewer, P.R., Ley, B.L., 2013. Whose science do you believe? explaining trust in sources of scientific information about the environment. Science Communication 35, 115.

Buchanan, M., 2012. It's a (stylized) fact! Nature Physics 8, 3.

Budescu, D.V., Broomell, S., Por, H.H., 2009. Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. Psychological Science 20, 299–308.

Burton, R.M., 2003. Computational laboratories for organization science: Questions, validity and docking. Computational & Mathematical Organization Theory 9, 91–108.

Carolan, M.S., 2008. The bright- and blind-spots of science: Why objective knowledge is not enough to resolve environmental controversies. Critical Sociology 34, 725–740.

Cash, D.W., Clark, W.C., Alcock, F., Dickson, N.M., Eckley, N., Guston, D.H., Jäger, J., Mitchell, R.B., 2003. Knowledge systems for sustainable development. Proceedings of the National Academy of Sciences 100, 8086–8091.

Celia, M.A., Gray, W.G., Hassanizadeh, S.M., Carrera, J., 1992. Validation of geo-hydrological models: Part i. Adv. Water Resour. 15, 1–274.

Covey, C., AchutaRao, K.M., Cubasch, U., Jones, P., Lambert, S.J., Mann, M.E., Phillips, T.J., Taylor, K.E., 2003. An overview of results from the coupled model intercomparison project. Global and Planetary Change 37, 103–133.

EC COM, 2011. Energy Road map 2050. Technical Report. COM(2011) 885 final, available at http://ec.europa.eu/energy/energy2020/roadmap/index_en.htm.

Exxon Mobil, 2012. The Outlook to Energy: A view of 2040. Technical Report. Retrieved from: www.exxonmobil.com/Corporate/energy_outlook_view.aspx.

Fisher, M.S., 2007. Software Verification and Validation - An Engineering and Scientific Approach. Springer, New York.

Funtowicz, S.O., Ravetz, J.R., 1990. Uncertainty and Quality in Science for Policy. Kluwer Academic Publishers, Dordrecht.

GCAM, 2012. The GCAM model. Technical Report. Documentation available at www.globalchange.umd.edu/models/gcam/.

GEA, 2012. Global Energy Assessment - Toward a Sustainable Future. Cambridge University Press. Cambridge UK and New York, NY, USA and the International Institute for Applied Systems Analysis, Laxenburg, Austria.

Gruebler, A., 2004. Transitions in energy use. Encyclopedia of Energy 6, 163–177.

Heston, A., Summers, R., Aten, B., 2006. Penn World Table Version 6.3. Technical Report. Center for International Comparisons of Production, Income and Prices at the University of Pennsylvania.

Hourcade, J.C., Jaccard, M., Bataille, C., Ghersi, F., 2006. Hybrid modeling: New answers to old challenges. The Energy Journal 2, Special issue, 1–12.

Huntington, H.G., Weyant, J.P., Sweeney, J.L., 1982. Modeling for insights, not numbers: the experiences of the energy modeling forum. Omega 10, 449–462.

Institute of Electrical Engineers Japan, 2011. Publication on World Energy. Technical Report. retrieved from http://eneken.ieej.or.jp/en/.

International Energy Agency, 2011. World energy outlook 2011. OECD Publishing.

IPCC, 2011. IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation. Prepared by Working Group III of the Intergovernmental Panel on Climate Change [Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Seyboth, K., Matschoss, P., Kadner, S., Zwickel, T., Eickemeier, P., Hansen, G., Schlömer, S., von Stechow, C. (eds)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1075 pp.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling & Software 21, 602–614.

Jasanoff, S., 2010. Testing time for climate science. Science 328, 695–696.

Jones, C., Romer, P., 2010. The new Kaldor facts: Ideas, institutions, population, and human capital. American Economic Journal: Macroeconomics 2, 224–245.

Kaldor, N., 1961. Capital accumulation and economic growth, in: Lutz, F., Hague, D. (Eds.), The Theory of Capital. St. Martins Press, London, pp. 177–222.

Kaya, Y., 1990. Impact of Carbon Dioxide Emission Control on GNP Growth: Interpretation of Proposed Scenarios. Technical Report. Paper presented to the IPCC Energy and Industry Subgroup, Response Strategies Working Group, Paris, mimeo.

Kloprogge, P., van der Sluijs, J.P., Petersen, A.C., 2011. A method for the analysis of assumptions in model-based environmental assessments. Environmental Modelling & Software 26, 289–301.

Konikow, L.F., Bredehoeft, J.D., 1992. Ground-water models cannot be validated. Advances in Water Resources 15, 75–83.

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.

Landry, M., Oral, M., 1993. In search of a valid view of model validation for operations research. European Journal of Operational Research 66, 161–167.

Macknick, J., 2011. Energy and CO2 emission data uncertainties. Carbon Management 2, 189–205.

Nordhaus, W.D., 2007. A review of the "stern review on the economics of climate change". Journal of Economic Literature 45, 686–702.

Nordhaus, W.D., 2012. The DICE model. Technical Report. Documentation at www.econ.yale.edu/ nordhaus/homepage/DICE2007.htm.

Oberkampf, W.L., Roy, C.J., 2010. Verification and Validation in Scientific Computing. Cambridge University Press, Cambridge.

Oderwald, R.G., Hans, R.P., 1993. Corroborating models with model properties. Forest Ecology and Management 62, 271–283.

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

Parker, P., Letcher, R., Jakeman, A., Beck, M.B., Harris, G., Argent, R.M., Hare, M., Pahl-Wostl, C., Voinov, A., Janssen, M., Sullivan, P., Scoccimarro, M., Friend, A., Sonnenshein, M., Barker, D., Matejicek, L., Odulaja, D., Deadman, P., Lim, K., Larocque, G., Tarikhi, P., Fletcher, C., Put, A., Maxwell, T., Charles, A., Breeze, H., Nakatani, N., Mudgal, S., Naito, W., Osidele, O., Eriksson, I., Kautsky, U., Kautsky, E., Naeslund, B., Kumblad, L., Park, R., Maltagliati, S., Girardin, P., Rizzoli, A., Mauriello, D., Hoch, R., Pelletier, D., Reilly, J., Olafsdottir, R., Bin, S., 2002. Progress in integrated assessment and modelling. Environmental Modelling & Software 17, 209–217.

Ravetz, R., 1997. Integrated Environmental Assessment Forum: developing guidelines for "good practice". Technical Report. ULYSSES WP-97-1. Available at www.jvds.nl/ulysses/eWP97-1.pdf.

Renn, O., Levine, D., 1991. Credibility and trust in risk communication, in: Kasperson, R.E., Stallen, P.J.M. (Eds.), Communicating Risks to the public: International perspectives.. Kluwer Academic Publishers, Dordrecht., pp. 175–218.

Risbey, J., Kandlikar, M., Patwardhan, A., 1996. Assessing integrated assessments. Climatic Change 34, 369–395.

Risbey, J., van der Sluijs, J., Ravetz, J., Janssen, P., 2001. A Checklist for Quality Assistance in Environmental Modelling. Technical Report. Dept. of Science, Technology and Society. Utrecht University. NWS-E-2001-11, ISBN 90-73958-65-2. Available at www.nusap.net/sections.php?op=viewarticle&artid=15.

Rosenberg, A., 1994. Instrumental Biology or The Disunity of Science. University of Chicago Press, Chicago.

Rykiel, E.J., 1996. Testing ecological models: the meaning of validation. Ecological Modelling 90, 229–244.

Saltelli, A., D'Hombres, B., 2010. Sensitivity analysis didn't help. A practitioner's critique of the Stern review. Global Environmental Change 20, 298–302.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S., 2008. Global Sensitivity Analysis, The Primer. John Wiley and Sons, New York.

Sargent, R.G., 2003. Verification and validation: verification and validation of simulation models., in: In Chick, S. , Sanchez, P. J., Ferrin, D., Morris, D. J. (eds.). Proceedings of the 2003 Winter Simulation Conference., pp. 37–48.

Sargent, R.G., 2010. Verification and validation of simulation models., in: In Johansson, B., Jain, S., Montoya-Torres, J., Hugan, J., Yucesan, E. (eds.). Proceedings of the 2010 Winter Simulation Conference., pp. 166–183.

Saysel, A., Barlas, Y., 2006. Model simplification and validation with indirect structure validity tests. System Dynamics Review 22, 241–262.

Schneider, S.H., 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.

van der Sluijs, J.P., 2002. A way out of the credibility crisis of models used in integrated environmental assessment. Futures 34, 133–146.

van der Sluijs, J.P., Craye, M., Funtowicz, F., Kloprogge, P., Ravetz, R., Risbey, R., 2005. Combining quantitative and qualitative measures of uncertainty in model-based environmental assessment: The NUSAP system. Risk Analysis 25, 481–492.

Smil, V., 2000. Perils of long-range energy forecasting: Reflections on looking far ahead. Technological Forecasting and Social Change 65, 251–264.

Sobol', I., 1993. Sensitivity analysis for non-linear mathematical models. Mathematical Modeling and Computational Experiment 1, 407–414.

Sobol', I., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation 55, 271–280.

Stern, N., 2007. The economics of climate change: The Stern review. Cambridge University Press, Cambridge and New York.

Sweeney, J.L., 1983. Energy model comparison: An overview., in: Thrall, R. (Ed.), Large Scale Energy Models - Prospects and Potentials. Westview Press, Boulder, pp. 191–217.

Tol, R.S.J., 2006. The Stern review of the economics of climate change: a comment. Energy & Environment 17, 977–981.

Voinov, A., Bousquet, F., 2010. Modelling with stakeholders. Environmental Modelling & Software 25, 1268–1281.

Weitzman, M.L., 2007. A review of "the Stern review on the economics of climate change". Journal of Economic Literature 45, 703–724.

Weyant, J.P., 2009. A perspective on integrated assessment. Climatic Change 95, 317–323.

White, D.D., Wutich, A.D., Larson, K.L., Gober, P.L., Lant, T.L., Senneville, C.L., 2010. Credibility, salience, and legitimacy of boundary objects: Water managers' assessment of a simulation model in an immersive decision theater. Science and Public Policy 37, 219–232.

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 , published online DOI 10.1007/s10584–012–0618–y.