International Congress on Environmental
Modelling and Software

1st International Congress on Environmental
Modelling and Software - Lugano, Switzerland -
June 2002

Jul 1st, 12:00 AM

Application of a Checklist for Quality Assistance in Environmental Modelling to an Energy Model

James Risbey

Jeroen Van der Sluijs

Penny Kloprogge

Jerry Ravetz

Silvio Funtowicz

See next page for additional authors

Presenter/Author Information

James Risbey, Jeroen Van der Sluijs, Penny Kloprogge, Jerry Ravetz, Silvio Funtowicz, and S.C. Quintana

Application of a Checklist for Quality Assistance in Environmental Modelling to an Energy Model

James Risbey[a]**, Jeroen van der Sluijs**[a]**, Penny Kloprogge**[a]
Jerry Ravetz[b]**, Silvio Funtowicz**[c]**, Serafin Corral Quintana**[c]
[a]Copernicus Institute for Sustainable Development and Innovation, Universiteit Utrecht, The Netherlands
[b]Research Methods Consultancy, London, UK
[c]Institute for the Protection and the Security of the Citizen, JRC, Italy

Abstract: Large, complex energy models present considerable challenges to develop and test. Uncertainty assessments of such models provide only partial guidance on the quality of the results. We have developed a model quality assistance checklist to aid in this purpose. The checklist is applied to an energy model for the problem of assessing energy use and greenhouse gas emissions. Use of the checklist suggests that results on this issue are contingent on a number of assumptions that are highly value-laden. When these assumptions are held fixed, the model is deemed capable of producing moderately robust results of relevance to climate policy over the longer term. Checklist responses also indicate that a number of details critical to policy choices or outcomes on this issue are not captured in the model, and model results should therefore be supplemented with alternative analyses.

Keywords: quality assistance, modelling

1 INTRODUCTION

Environmental models are often used in policy assessment exercises. Yet, because of their size and complexity, it is difficult to know how much trust should be placed in the results. To cope with this, there has been a lot of effort to characterize the uncertainties associated with the models and their projections van der Sluijs [1997]. However, uncertainty estimates alone are necessarily incomplete on models of such complexity and provide only partial guidance on the quality of the results. The conventional method to ensure quality in modelling domains is via model validation against observed outcomes. Unfortunately, the data are simply not available to carry out rigorous evaluations of many models Risbey et al. [1996]; Beck [2002]. Without the ability to validate the models directly or perform comprehensive uncertainty analyses, other forms of quality assessment must be utilized. Indeed, evaluation of models is increasingly being cast in broader terms to encompass issues of purpose and use (as well as performance) and quality assurance in design of tools and controlling procedures Beck [2002]. This work follows that broader conception in including each of these elements in a checklist format to be used in aiding the modelling process. The model checklist is also situated in a broader assessment context that strives for

greater transparency, accountability, effectiveness, and a democratizing of expertise in assessment processes European Commission [2001].

For complex coupled models there are many pitfalls in the modelling process and some form of rigour is essential to yield quality. Thus, a modeller has to be a good craftsperson Ravetz [1971, 1999]. Discipline is maintained by controlling the introduction of assumptions into the model and maintaining 'good practice'. What is needed in this case is a form of heuristic that encourages self-evaluative systematization and reflexivity on pitfalls. The method of systematization should not only provide some guide to how the modellers are doing; it should also provide some diagnostic help as to where problems may occur and why. We have developed a model quality assistance checklist for this purpose, which is available via the web at `http://www.nusap.net`.

The philosophy underlying the checklist is that there is no single metric for assessing model performance and that, for most intents and purposes, there is no such thing as a 'correct' model. Rather, models need to be assessed in relation to particular functions. Further, that assessment is ultimately about quality, where quality relates a process/product (in this case a model) to a given function. The point is not that a model can be classified as 'good' or 'bad', but that there are 'better' and 'worse' forms

of modelling practice, and that models are 'more' or 'less' useful when applied to a particular problem. The checklist is thus intended to help guard against poor practice and to focus modelling on the utility of results for a particular problem. That is, it should provide some insurance against pitfalls in process and irrelevance in application. The questions in the checklist are designed to uncover at least some of the more common pitfalls in modelling practice and application of model results in policy contexts. The output from the checklist is both indirect, via reflections from the modeller's self assessment, and direct in the form of a set of potential pitfalls triggered on the basis of the modeller's responses.

The checklist is structured as follows. First there is a set of questions to probe whether quality assistance is likely to be relevant to the intended application. If quality is not at stake, a checklist such as this one serves little purpose. The next section of the checklist aims to set the context for use of the checklist by describing the model, the problem that it is addressing, and some of the issues at stake in the broader policy setting for this problem. The checklist then addresses 'internal' quality issues, which refers to the processes for developing, testing, and running the model practiced within the modelling group. A section on 'users' addresses the interface between the modelling group and outside users of the model. This section examines issues such as the match between the production of information from the model and the requirements of the users for that information. A section on 'use in policy' addresses issues that arise in translating model results to the broader policy domain, including the incorporation of different stakeholder groups into the discussion of these results. The final section provides an overall assessment of quality issues from use of the checklist and provides feedback in the form of a set of potential pitfalls for the model application.

In what follows we describe the application of the checklist to an energy model for the purpose of estimating greenhouse gas emissions under the IPCC SRES B1 energy scenario Nakicenovic et al. [2000]. The energy model in question is the TIMER model De Vries et al. [2000, 2002]; van Vuuren and de Vries [2001]. TIMER is the energy model component of the IMAGE 2 integrated assessment model Alcamo et al. [1998].

2 APPLICATION TO TIMER

The application of the checklist to the TIMER model was carried out in an interview with mod-

eller, Detlef van Vuuren, by Risbey and van der Sluijs.

2.1 Use of the checklist

The first questions in the interview were aimed at assessing the utility of the checklist for the given application. These showed that there is some question as to the accuracy of model results, some interpretation and judgement of results is required, and that the public is concerned about process and results regarding the model application. Thus, quality considerations seem relevant to this application and use of the checklist is warranted.

2.2 Problem context

The problem addressed by the TIMER model for this application is how will greenhouse gas emissions develop given different world views and assumptions about population and economic growth (as specified in the SRES scenarios)? Model output variables of relevance to this problem are primary energy production and consumption, final energy consumption, and biomass production. Responses to questions in the checklist focus on these variables unless otherwise indicated.

For the application of the model described above, the intended users identified are the IMAGE group, the energy modelling community, and national and international policymakers and stakeholders concerned about climate change. A number of groups were identified as having particular interests in the outcome of the research on this problem. Such interest was apparent in earlier discussions on the SRES scenarios within the IPCC. For example, one could imagine that the Middle East oil producing regions favour scenarios that imply that fossil fuel use is benign for the climate, and to some degree they tried to influence the shaping and selection of the SRES scenarios to this end. Other participants argued for setting high emission baselines in the SRES scenarios to demonstrate the need for climate policies. After publication of SRES, it became clear that some countries and NGO's are skeptical of the B1 SRES scenario as they fear that it could be interpreted to undermine the need for active climate policies. In short, the stakes for the research are relatively high and a number of different groups have vested interests in the outcome.

The research of the IMAGE/TIMER group is funded via the Dutch environment ministry. The views on climate policy of members of the ministry

are of course known to the modelling group. Some model results over time were assessed to be convergent with these views and some not. In other words, no systematic bias to funder views was assessed.

2.3 Values and key parameter identification

Value choices are often key determinants of outcomes in energy modelling contexts Kloprogge and van der Sluijs [2002]. A long list of key value-laden issues were identified of relevance to the model application. Starting with the SRES scenarios, values enter into the characterization of 'globalized' versus 'regionalized' worlds. Indirectly, the SRES scenarios seem to embody an assumption that globalization is 'good' for the environment. This assumption is operationalized via assumptions about the different economic growth rates between globalized and regionalized worlds and via those on the demographic transition, whereby increases in GDP are assumed to automatically lead to reductions in birthrates. This leads to lower emissions in the 'globalization' scenarios. Another interesting point is that final energy consumption was specified in the SRES scenarios as a 'harmonized' parameter. This means that the other models were more or less constrained to adopt the assumptions on globalization for instance made by the SRES 'marker' models.

In the TIMER model framework further values-related issues identified were the learning rates for technology development in the energy sector, structural change in the energy-economic system, trade contraints, the availability of resources, technological development in energy consumption and efficiency, and payback times for investments. On the century long time scale, the model was assessed to be substantially conditioned by value issues. The modeller's assessment of value-ladeness is consistent with those made in a recent workshop on TIMER in which participants used a NUSAP pedigree matrix Funtowicz and Ravetz [1990] to score the value-ladeness of model variables van der Sluijs et al. [2002].

Most of the key parameters governing spread in model output variables of interest for this problem have been identified through sensitivity studies van der Sluijs et al. [2002]. They include population and GDP (from the SRES scenario), structural change in the economy, learning factors for energy systems, available resources, and investment payback times. Note that there is considerable overlap between the list of key variables governing spread in output and the list of key value-laden variables.

2.4 Model structure and validation

Various alternatives for model structure were identified in the checklist interview. In particular, some models take a 'bottom-up' approach to modelling the energy system from the component technologies and sectoral demands. Such models provide good resolution of the energy system but typically do not include feedbacks between the energy and economic systems. Other models pursue a 'top-down' approach from macroeconomic considerations. These models do include feedbacks between the energy and economic systems, but typically provide little resolution of the energy system. The TIMER model is by choice somewhere in between and contains characteristics of both types of energy models. In particular, it shares some of the assumptions of bottom-up models. The effects of alternative model structures have not been tested explicitly. Implicit testing is carried out by comparison of results with other energy models. Results for the key output variables were judged to be at least moderately sensitive to the structural underpinnings of the model.

Validation of the model has been carried out on the limited data available and indirectly via model intercomparison (particularly via the SRES process). Validation has been aided by the fact that much of the available data is at the same level of aggregation as the model, but this data is quite uncertain in some regions.

2.5 Robustness and accuracy of results

Model results for final energy consumption were judged to be moderately robust in that they could probably be changed by a factor of two or so without much tinkering with parameter values, but not by a factor of ten without requiring implausible changes to the model. For a hypothetical sensitivity study encompassing most of the major assumptions, the resulting spread in energy consumption was assessed to be less than a factor of two given the B1 scenario, but larger than that when encompassing the full set of SRES assumptions on population and growth. In translating energy consumption to CO_2 emissions, the level of accuracy assessed for CO_2 emissions was judged to be around 10% given the assumptions of the B1 scenario.

The modeller's assessment of the levels of accuracy required for model results to be useful in the policy process was to better than 10% for short term (2-3 decades) energy planning, but much less ac-

curacy than that for long term (century scale) climate policy such as entailed in the Kyoto protocol. Given the levels of accuracy assessed for model outputs, model results were deemed to be too coarse for short term planning, but of about the required level of accuracy for assessing the greenhouse gas implications of long term scenarios such as B1. On the question of whether the model provides useful answers for climate policy assessment, the modeller differentiated between assumptions at the SRES and B1 level. He noted that the SRES scenarios depend in part on one's world view and it is difficult to differentiate among them on the basis of plausibility. Thus, when encompassing assumptions at the SRES level related to population, trade, and growth, model results were deemed to be relevant to the policy process, but with unknown plausibility. With these factors held fixed for the B1 scenario, model results were judged to be 'relevant and plausible'.

2.6 Model role in policy

The modeller was asked what role the model *should* play in setting policy on this issue. He replied that any particular energy model should provide only a weak guide to policy, but that the class of energy models taken together could provide a more general guide for policy. This response was consistent with his assessment of how models actually *are* used in the policy process. He noted that models are used rhetorically, pro or con particular policies and for community building. He noted that the SRES process helped communicate the notion of different possibilities and worlds between modellers and policymakers. The modeller provided an example of why model results are best used in combination than alone for policy. On the question of whether to delay action to mitigate greenhouse gas emissions or act now, he developed a list of six reasons for each position (twelve total). He noted that three of these arguments could be addressed in one energy model and three in another. On a more cautionary note, he noted that six of the twelve arguments were not addressed in any of the models he considered. This is consistent with his response on how models ought to be used, which stopped short of the category specifying that 'policies should be directly keyed to specific model results'.

2.7 Model development

Questions aimed at model development practices indicated that there has not been a systematic process for evaluating model assumptions, nor have the effects of increases in model complexity been moni-

tored by systematic routines. To be sure, this is currently normal practice for the field. Some attention is given to model anomalies (results departing from expectations based on theory, data, or other models) and discussed in the broader modelling community. One difficulty, if not necessarily an anomaly, in the model is the need to calculate certain quantities as functions of price rather than amount. This is a constraint based on available data. An anomaly in the sense of differences with other models is the assumption of saturation of energy demand in the formula for structural change. This results in TIMER being at the low end of the range of energy demand calculated by the SRES group of models Nakicenovic et al. [2000]. However, this is a consequence of a conscious choice on how to model energy demand rather than an unusual outcome per se. Unresolved anomalies and assumptions such as the above were assessed to be treated openly in relation to both users and the public.

2.8 Model access

Questions on the access of outsiders to the IMAGE/TIMER models indicated mixed results. At present there is an effective monopoly of access to the model. The model is in the process of being documented De Vries et al. [2002], the source code is public (upon request), and other groups do use the model. However, these groups require assistance to use the model, which is fairly complicated. Specialized software (the M compiler) is needed to change the model, though hardware is typically not a constraint because the model is not computationally demanding. With regard to the broader policy and stakeholder community, there has been minimal inspection or use of the model, which is more or less typical for energy models. The presence of value judgements in the model is communicated to policy audiences, though such audiences are typically only partially aware of the implications of the different value choices for model results.

2.9 Overall assessment

The modeller's overall assessment for the problem of projecting energy consumption and greenhouse gas emissions is that model results can be used with 'caution' (on a scale from 'extreme caution' to 'caution' to 'confidence' to 'high confidence'). His broad reasoning is that the different energy models can be useful if used in conjunction, but that they do not include all pertinent factors. For example, he noted that there are more reasons for energy scenarios to diverge based on factors not included in the

models than based on factors that are captured in the models.

2.10 Pitfalls

The following list of potential 'pitfalls' were generated in response to the TIMER checklist run. The list of pitfalls is generated via a preset algorithm on the basis of checks of the responses coded for each of the questions. The algorithm checks for inconsistencies among responses and for responses that indicate potentially poor or inappropriate practice. The results generated from this step were then checked in consultation with the modeller. Some consultation on results is useful because it is difficult to generalize pitfalls. That is because there are not always single 'best' answers to the questions. What constitutes good practice in one domain may be in conflict with the requirements of good practice in another, and the resolution of such conflicts will often depend on the context. Thus, the list of pitfalls should be viewed as a guide only:

- Uncertainty in input values is only partially represented by the sensitivity runs carried out to date. Thus, the list of key parameters selected for this problem is not necessarily complete.

- Since uncertainties have not been propagated through the model from inputs to outputs, one can not rigorously state what the final error bars are. It is important to be cautious of this fact in interpreting model results.

- Since alternative model structures have not been tested and have only indirectly been addressed through model intercomparison, the effects of structural uncertainty are partly unknown. More effort may need to be devoted to exploring effects of alternative model structures.

- Model results are sensitive to uncertainty in model structure formulation. This fact should be noted when presenting results.

- The key results are potentially very sensitive to uncertainty in parameter values. The non robust nature of the energy system represented by the model should be signalled to users.

- There is a broad spread of possible output values in key model results. Some of the uncertainty may be irreducible, and high spread does not necessarily imply low quality.

Nonetheless, the results should be checked against users needs to determine if the spread is narrow enough to be useful.

- There is a lack of systematic processes for managing development of the model.

- It is difficult for outside groups to run the model because of specialized requirements of software and familiarity with a large, complex body of code. This means that model results are effectively not very reproducible by outsiders, increasing the likelihood of error and decreasing general acceptance of the results.

- The model could benefit from more involvement of stakeholders in using or inspecting the model. The reasons for relatively low stakeholder involvement should be ascertained if not already known.

- Users of model results in policy are at best partially aware of the implications of different value choices in the model. Better communication seems warranted in this regard.

3 CONCLUSIONS

The list of potential pitfalls generated for the TIMER run through the checklist are intended to apply to use of TIMER results on energy scenarios and greenhouse gas emissions. It is clear from use of the checklist that results on this issue are contingent on a number of assumptions that are highly value-laden. When these assumptions are held fixed, the model is deemed capable of producing moderately robust results of relevance to climate policy over the longer term. However, it is critical that the effects of value choices be communicated as clearly as possible in assessing model results. Checklist responses also indicate that a number of details critical to policy choices or outcomes on this issue are not captured in the model, and model results should therefore be supplemented with alternative analyses.

While these comments are made in reference to testing of the checklist on TIMER, they would apply broadly to other energy models as well. That is because other energy models must make the same assumptions and compromises as TIMER in approaching this problem. They may make different choices in how best to do this, but that does not weaken the force of many of the most critical assumptions or reduce the inherent value-loading of the analysis.

4 ACKNOWLEDGMENTS

REFERENCES

Alcamo, J., namesleft>1 Leemans, R.,numnames>2 and lastname Kreileman, E., editors. *Global change scenarios for the 21st century. Results from the IMAGE 2.1 model.* Elsevier Science, London, 1998. 572pp.

Beck, B. *Model evaluation and performance. In: Encyclopedia of Environmetrics*, pages 1275–1279. John Wiley & Sons, New York, 2002. Volume 3.

De Vries, B., J. Bollen, A. Bouwman, M. den Elzen, M. Janssen, and E.Kreileman. Greenhouse gas emissions in an equity-, environment- and service-oriented world: an IMAGE-based scenario for the 21st century. *Tech. Forecasting and Social Change*, 63(2-3):137–174, 2000.

De Vries, B., D. van Vuuren, M. den Elzen, and M. Janssen. The TARGETS-IMAGE energy regional model (TIMER): Technical documentation. Technical report, National Institute for Public Health and the Environment, Bilthoven, NL, 2002. Report 481508014.

European Commission. White paper on governance. Report of the working group: Democratizing expertise and establishing scientific reference systems (Group 1b). Technical report, European Commission, Brussels, 2001. 26pp.

Funtowicz, S. and J. Ravetz. *Uncertainty and Quality in Science for Policy.* Kluwer, Dordrecht, 1990. 229pp.

Kloprogge, P. and J. van der Sluijs. Choice processes in modelling for policy support. In *Proceedings of the International Environmental Modelling and Software Society*, pages 1–6, Lugano, Jun 2002. IEMSS.

Nakicenovic, N., J. Alcamo, G. Davis, B. de Vries, and 24 others. *Special Report on Emissions Scenarios: A Special Report of the Intergovernmental Panel on Climate Change.* Cambridge Univ. Press, Cambridge, UK, 2000. 599pp.

Ravetz, J. *Scientific knowledge and its social problems*. Clarendon Press, Oxford, 1971. Reprint: Transaction, New Brunswick NJ, 1996, 449pp.

Ravetz, J. Developing principles of good practice in integrated environmental assessment. *Int. J. Env. and Pollution*, 11(3):243–265, 1999.

Risbey, J., M. Kandlikar, and A. Patwardhan. Assessing integrated assessments. *Clim. Change*, 34 (3-4):369–395, 1996.

van der Sluijs, J. *Anchoring amid uncertainty. On the management of uncertainties in risk assessment of anthropogenic climate change.* Universiteit Utrecht, Utrecht, 1997. 260pp.

van der Sluijs, J., J. Risbey, S. Corral Quintana, and J. Ravetz. Uncertainty assessment of the TIMER model: Using the NUSAP method. In *Proceedings of the International Environmental Modelling and Software Society*, pages 1–6, Lugano, Jun 2002. IEMSS.

van Vuuren, D. and B. de Vries. Mitigaton scenarios in a world oriented at sustainable development: the role of technology, efficiency and timing. *Clim. Policy*, 1(2):189–210, 2001.