Environmental Modeling and Assessment (2005) 10:63–79
DOI 10.1007/s10666-004-4267-z

Application of a checklist for quality assistance in environmental modelling to an energy model

James Risbey [a], Jeroen van der Sluijs [b], Penny Kloprogge [b], Jerry Ravetz [c], Silvio Funtowicz [d] and Serafin Corral Quintana [e]

[a] *School of Mathematical Sciences, Monash University, 3800 Australia*
[b] *Copernicus Institute, Universiteit Utrecht, 3584 CS Utrecht, The Netherlands*
[c] *Research Methods Consultancy, 106 Defoe House, Barbican, London, UK*
[d] *Institute for the Protection and the Security of the Citizen, EC-JRC, 1-21020 Ispra, Italy*
[e] *Economia de las Instituciones, University of La Laguna, Spain*

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 model checklist provides diagnostic output in the form of a set of pitfalls for the model application. The checklist is applied here 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, climate, energy

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 results from the models. To cope with this, there has been considerable effort to characterize the uncertainties associated with the models and their projections [1–3]. 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 [4]. Unfortunately, the data are simply not available to carry out rigorous evaluations of many models [4–6].

Lack of validation data is critical in the case of complex models spanning human and natural systems because such models typically require: socio-economic data which has frequently not been collected; data related to value dimensions of problems that is hard to define and quantify; data on projections of technical change which must often be guessed at; data on aggregate parameters like energy efficiency which is difficult to measure and collect for all the relevant economies; geophysical data on fine spatial and temporal scales worldwide that is not generally available; data pertinent to non-marginal changes in socio-economic systems which is difficult to collect; and experience and data pertaining to system changes of the kind simulated in the models for which we have little precedent or access.

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 [6]. 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 [7].

For complex environmental models there are many pitfalls in the modelling process and some form of rigour is essential to yield quality [5]. Complex models are large collections of software and are prone to all the standard problems associated with software development [8]. When model code reaches a critical size (exceeded by all but the simplest models), error checking tends to occur only (or mostly) when results depart from expectations, and thus has more of the character of systematic bias than systematic checking. If the model is scrutinized mostly in those situations when strange results occur, then strange results will tend to be corrected (when a model error is found) so that the model reconfirms expectations. On the flip side, when model results confirm expectations, error checking tends to be only cursory and the model will go uncorrected, even if wrong. Thus, there is a bias towards confirmation of expectations. We call this bias 'WYGIWYE' – What You Get Is What You Expect. In order to guard against WYGIWYE (and other biases [9]) a modeller has to be a good craftsperson [10]. Discipline is maintained by controlling the introduction of assumptions into the model and maintaining 'good practice' [11]. 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 described here and included as appendix.

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 (section A.2). If quality is not at stake, a checklist such as this one serves little purpose. The checklist is fairly comprehensive, and many modellers will not have the time or need to complete the entire checklist. For that reason, the checklist contains a set of screening questions at the front to allow the modeller to identify the parts of the checklist that are potentially most useful for their application. 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 (section A.3). The checklist then addresses 'internal' quality issues, which refers to the processes for developing, testing, and running the model practiced within the modelling group (section A.4). A section on 'users' addresses the interface between the modelling group and outside users of the model (section A.5). 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 (section A.6). The final section of the checklist 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 (section A.7).

We introduce the checklist here by describing an application of the checklist to an energy model. Since the checklist does not attempt to grade models per se, but relates to their fitness for given functions, we assess use of the energy model for the purpose of estimating greenhouse gas emissions under the IPCC SRES Bl energy scenario [12]. The energy model in question is the TIMER model [13–15]. TIMER is the energy model component of the IMAGE 2 integrated assessment model [16]. The IMAGE model constitutes the broader integrated model framework linking social and natural systems in which TIMER is embedded.

The remainder of the paper is organized as follows. First, we provide a brief description of the TIMER model. Then we outline the results of a self-assessment of the TIMER model for the purpose of projecting greenhouse gas emissions, guided by the checklist. Finally, we provide a diagnosis of potential pitfalls in using the model for this purpose as highlighted by the checklist assessment. We intend this assessment to shed light on both the utility of the checklist and on key quality issues at stake in energy modelling.

2. Description of TIMER

The TIMER model is used to analyse and project the long term dynamics of the world's energy system. It comprises a set of submodels describing energy demand, supply, and prices. The demand submodel represents demand for fuel and electricity in five sectors (industrial, transport, residential, services, and other sectors) for a range of different energy carriers. The demand for energy services is driven by changes in population and economic activity. Energy supply is represented in a series of submodels for solid, liquid, and gaseous fuels, and takes into account the primary commercial fossil and non-fossil sources. The model is broken down into seventeen world regions. It is calibrated to reproduce the major world energy trends in the 1975–1995 period, and is run forward in time to the year 2100. TIMER simulates year-to-year investment decisions based on a combination of bottom-up engineering information and specific rules on investment behaviour, fuel substitution, and technology. This is in contrast to optimization-based models which optimize decisions over the complete modelling period on the basis of perfect foresight.

The main model inputs to TIMER are regional population and macro-economic activity levels, along with assumptions on rates of technological development and resource availability. The main model outputs are the temporal evolution of energy demand, fuel costs, and competing supply technologies in the different model regions. These outputs in turn are fed to an emissions submodel to calculate emissions of greenhouse gases, ozone precursors, and acidifying compounds. The inputs for this case study of model projections of greenhouse emissions are shaped by the IPCC SRES Bl scenario governing population and energy development over the 21st century [12]. We assess uncertainties in TIMER outputs, both assuming the IPCC scenario as given (with no uncertainty), and taking account of uncertainty in the underlying IPCC scenario inputs. Readers interested in further details of the TIMER model are referred to reference [14].

3. Application of the checklist to TIMER

The application of the checklist to the TIMER model was carried out in an extended interview with TIMER modeller, Detlef van Vuuren, by Risbey and van der Sluijs. The responses were shaped by dialogue with van Vuuren, but the following descriptions represent our interpretations of that dialogue.

3.1. Use of the checklist

The first questions in the interview were aimed at quickly assessing the relevance and utility of the checklist for the given application for assessing long term greenhouse gas emissions. These are cast in the checklist as a set of screening questions to determine whether quality issues are really at stake in use of the model. There is no point completing a detailed checklist for quality assistance if quality concerns are not relevant to the issue in question. Responses to the screening questions 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. The screening section of the checklist also serves to quickly isolate the potentially most cricital areas for quality assistance so that users with limited time can be directed straight to the relevant parts of the checklist.

3.2. 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 modelling 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 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 Bl 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.

3.3. Values and key parameter identification

Value choices are often key determinants of outcomes in energy modelling contexts [1,17–19]. 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 workshop on TIMER in which participants used a NUSAP pedigree matrix [2] to score the value-ladeness of model variables [20].

Most of the key parameters governing spread in model output variables of interest for this problem have been identified through sensitivity studies [20]. 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.

3.4. 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.

3.5. 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 Bl 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 Bl 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 accuracy than that for long term (century scale) climate policy such as entailed in the Kyoto protocol [21]. 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 Bl. On the question of whether the model provides useful answers for climate policy assessment, the modeller differentiated between assumptions at the SRES and Bl 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 Bl scenario, model results were judged to be ‘relevant and plausible’.

3.6. 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’.

3.7. 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 monitored 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 [12]. 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.

3.8. 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 [14], 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.

3.9. 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.

3.10. Diagnosis of pitfalls

The main tangible output from use of the checklist is a diagnosis of potential pitfalls in applying the model to the given problem. 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 cannot 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.

4. 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.

The checklist could be used at various stages in the development of a model and application to a particular problem. In the energy model example given here the checklist was employed after the initial development of the model and during the ongoing application of projecting greenhouse gas emissions. The diagnosis of pitfalls can help in further model development and in effectively connecting the model to the policy process – by avoiding the more obvious pitfalls in this process. The checklist could also be used proactively prior to development of a model to shape the process of model development itself. However, it should also be kept in mind that the checklist is oriented at the role of models only. It does not provide assistance for use with the various other tools and aspects that can be included in the environmental analysis process (e.g. [22]).

Finally, we hope that we have demonstrated that a checklist such as the one developed has potential to provide useful diagnostic aid in the quality assessment process for complex environmental models.

Acknowledgements

This work was assisted by Detlef van Vuuren and Bert de Vries, as well as by colleagues at the University of Utrecht and RIVM, particularly Wim Turkenburg, Pita Verweij, Peter Janssen, and Arthur Petersen. Partial funding was provided by RIVM and the Australian Research Council. The checklist described here is available for online use via the web at http://www.nusap.net.

Appendix. A checklist for quality assistance in environmental modelling

A.1. Introduction

The goal of this checklist is to assist in the quality control process for environmental modelling. The point of the checklist is not that a model can be classified as ‘good’ or ‘bad’, but that there are ‘better’ and ‘worse’ forms of modelling *practice.* We believe that one should guard against poor practice because it is much more likely to produce poor or inappropriate model results. Further, model results are not ‘good’ or ‘bad’ in general (it is impossible to ‘validate’ a model in practice), but 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 insurance against pitfalls in process and irrelevance in application. The checklist is designed largely for internal use (within a modelling group) for self-assessment. It can be used as a self-elicitation by competent practitioners, to give form to their own judgements about the models they know intuitively. 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.

Before commencing the checklist, a few definitions are in order. For the purposes of this checklist we diffentiate between ‘users’ and ‘stakeholders’ as follows: A ‘user’ is someone who exercises the model or who uses its output in some application. A user is necessarily aware of the existence of the model. A stakeholder is one who either participates in the policy process regarding the issue at hand, or who is affected by that process in some way. Stakeholders may or may not be aware of the existence of the model (or of the policy process for that matter).

The checklist is arranged 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 checklist itself is fairly long, and many modellers will not have the time or need to complete the entire checklist. For that reason, we have provided a set of screening questions at the front to allow the modeller to identify the parts of the checklist that are potentially most useful for their application (section A.2). The first section of the checklist proper (section A.3) aims to set the context for use of the checklist by describing the model, the problem that it is addressing here, and some of the issues at stake in the broader policy setting for this problem. Section A.4 addresses ‘internal’ quality issues, which refers to the processes for developing, testing, and running the model practiced within the modelling group. Section A.5 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. Section A.6 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.

A.2. Screening questions

A.2.1. Should you use this checklist at all?

The checklist is designed for use on relatively complex models where validation of model outputs is not possible or is at best partial. In complex model domains the density of pitfalls is high and some form of rigour in the modelling process is needed to avoid them. The checklist is designed to help mark some of the more obvious pitfalls. If the model is well calibrated and validated by appropriate independent data then many of these pitfalls can be effectively avoided and the checklist may not be necessary. If the model itself is relatively simple and transparent in its use and assumptions then the pitfalls entailed are of a qualitatively different nature than those envisaged here and some other form of checklist might better be used.

Beyond these considerations, one should also be satisfied that quality is relevant to your application. This is not always the case. Sometimes quality is irrelevant because a model is widely accepted by all parties as an imperfect, but appropriate, metric on which to base decisions or gauge input to decisions. Quality may also be an irrelevant concern if the model is simply ignored by all. For quality to be at stake, the results of the model must be considered relevant by at least some stakeholders, and there must be some contention about the status of those results. The following questions are designed to help you decide whether quality is at stake in your application:

A.2.1.1. Is the model well validated by adequate empirical data?

some question as to the accuracy of results for this application	accuracy of results not in question for this application
□	□

A.2.1.2. Is the model simple enough that you can trace all model results to changes or responses of specific model variables?

some interpretation and judgement entailed in evaluating results	model results transparent and intuitive
□	□

A.2.1.3. Is the model well accepted for use on the desired application by:

peers	□
users	□
stakeholders	□

A.2.1.4. Is the model application salient to stakeholders and the public agenda?

model results widely ignored	model results sought by some	model results keenly sought by range of stakeholders
□	□	□

A.2.1.5. Is the legitimacy of the model community an issue among stakeholders?

community widely discredited	mixed acceptance	community widely accepted
□	□	□

A.2.1.6. Is public accountability of the science important to the policy process?

public concerned at most with the end results	public concerned with process and results	public focused on the process of the science
□	□	□

A.2.2. Which parts of the checklist are potentially useful?

Section A.3 should be completed in any run through the checklist since it sets the problem on which the checklist is being applied. Other sections or subsections of the checklist may not be germane for some models or model applications. The questions in this section are designed to help select sections that are likely to be more useful in highlighting relevant pitfalls.

A.2.2.1. Internal strength

Section A.4 relates to the maturity of model development and testing processes. Immature models or novel applications are more likely to benefit from this section. If the model and application are well established, consider skipping this section.

If not, circle the subsection numbers as appropriate to indicate that a section should be completed.

	Section to complete
If there has not been extensive sensitivity and parameter testing	4.1
If alternative model structures have not been explored	4.2
If the model is not extensively validated	4.3
If the model is sensitive to uncertainty in model parameters	4.4
If the model is not well documented or not widely used	4.5

A.2.2.2. Interface with users

Section A.5 helps assess whether the outputs from the model are appropriate and relevant to the needs of the user community. If there has been a long history of successful interaction with users, consider skipping this section. If not, circle the subsection numbers as appropriate to indicate that a section should be completed.

	Section to complete
If users have not been involved in the process of refining output variables and do not have well established procedures for incorporating them into their applications	5.1 and 5.2
If use of model data has been an issue for user applications	5.3, 5.4 and 5.5
If there have been problems with users misusing model results	5.6

A.2.2.3. Use in policy

Section A.6 examines the role of model results in shaping policy procedures or outcomes. If model results are widely accepted and generally uncontroversial for the application in question, consider skipping this section. If not, circle the section numbers as appropriate to indicate that a section should be completed.

	Section to complete
If stakeholders have not been involved in the process of model experiment design	6.1
If there is not an agreed format and means for using model results in policy	6.2
If stakeholders are not generally aware of the assumptions underlying the key model results	6.3

A.3. Model and problem domain

This section sets the context for use of the checklist by setting out what the problem is, what's at stake, how model output is relevant, and what role it will play in addressing the problem.

A.3.1. Model name:

Provide a brief genealogy of the model. Cite the main documents describing the model.

A.3.2. Intended function or application

A.3.2.1. Describe the problem being addressed

A.3.2.2. Describe the way in which the model will aid solution of the problem

A.3.2.3. List the most important model output variable (or set of variables) of relevance to this problem

Note that your responses to the checklist questions will often be framed in terms of these variables.

A.3.3. Intended users

Identify the users of model results and interested stakeholders.

A.3.4. Problem domain

A.3.4.1. For this problem, what are the key value issues?
List them and categorize them according to how central they are to this problem:

value	peripheral	relevant	central
	□	□	□
	□	□	□
	□	□	□
	□	□	□

A.3.4.2. List any pertinent facts that are in dispute?
A.3.4.3. Identify any groups vested in the outcome of research on this issue?
Briefly state the position favoured by each group if an identifiable position exists.

A.3.4.4. Who funds your groups research on this issue?
A.3.4.5. What role should models play in setting policy on this issue?

none	heuristic or weak guide	a general guide	policies directly keyed to specific model results
□	□	□	□

Explain.

A.4. Assessment of internal strength

This section is intended to examine practices within the modelling group and their relationship to quality issues.

A.4.1. Parametric uncertainty and sensitivity

A.4.1.1. Has the strength of the input data been assessed?

not tested	partially tested	well tested and used	peer reviewed
□	□	□	□

A.4.1.2. Have the key parameters (governing spread in model output) been identified?
List them.

A.4.1.3. How has uncertainty in key parameters been assessed?
How is uncertainty in model variables represented?

not at all	vary parameter values	using pdf's
□	□	□

A.4.1.4. Has a Monte Carlo or equivalent process been used for error propagation, and with what results?

not at all	propagation of errors indicates broad spread	propagation of errors indicates minimal spread
□	□	□

A.4.2. Structural uncertainty assessment

A.4.2.1. Are there plausible alternative model structures for representing the same empirical data or relations between variables?
Describe them.

A.4.2.2. If alternative structures were not tested, explain briefly why not

A.4.2.3. How do you expect results (for the key output variables indicated in section A.3.2) to vary when using different structures?

trivially	moderately	radically
□	□	□

A.4.2.4. Can differences among results (for key outputs) be explained in terms of specific model processes or changes?

black box view	some understanding	well understood
□	□	□

A.4.2.5. How was the system boundary defined?

Describe the forms the boundaries take and the reasons made for choices.

A.4.2.6. Have the consequences of alternative boundary choices been examined?

What are the implications for results (for key outputs)?

trivial	moderate	radical
□	□	□

A.4.2.7. Was uncertainty analysis built into the model with its initital design? If not, how was it instituted?

A.4.2.8. Were non-modelling approaches considered?

List any non-modelling approaches considered for addressing this problem and rank the relevance of each.

approach	peripheral	relevant	essential
	□	□	□
	□	□	□
	□	□	□

A.4.3. Validation

A.4.3.1. What kinds of model validation have been carried out?

Check all that apply:

◊ On independent data sets, avoiding calibration data.
◊ On partially independent data (some overlap with calibration data).
◊ By proxy (indirect) indicators.
◊ By model intercomparison.
◊ Other. Describe.
◊ None.

A.4.4. Robustness

A.4.4.1. How vulnerable is the model to "hack and crack"? (Is it possible to produce an arbitrarily chosen output by tweaking the system?)

If you were asked to change the main result of the model for this problem by a factor of 2, how much would you need to 'tweak' the most sensitive parameter values:

barely – well inside range of expert opinion	moderately – moving to tails of expert distributions	radically – outside expert disbtributions
□	□	□

If you were asked to change the main result of the model for this problem by a factor of 10, how much would you need to ‘tweak’ the most sensitive parameter values:

barely –	moderately –	radically –
well inside range	moving to tails of	outside expert
of expert opinion	expert distributions	disbtributions
□	□	□

A.4.4.2. Are the sets of assumptions related to model structure, boundary choice, and parameter values employed in experiment design wide enough to be credible?

Given your assessment of the critical assumptions, your experimental design has encompassed and tested:

few of the major	some of the major	most of the major
assumptions	assumptions	assumptions
□	□	□

A.4.4.3. Is the range of results narrow enough to be useful?

Provide your assessment (or estimate, if you did not check the rightmost box above) of the spread of model results (for key outputs) for a sensitivity study encompassing most of the major assumptions:

order of magnitude	a factor of 2	better than 10%
□	□	□

A.4.5. Model development practices

A.4.5.1. Has there been a systematic process for evaluating model assumptions, including their influence on the total structure and their possible pitfalls?

Describe the process.

A.4.5.2. Have the effects of increases of complexity in the model (including new processes) been monitored by systematic routines?

For example, do you perform a sensitivity analysis when the model is changed:

occasionally, focusing	occasionally, including	often, including
on a few parameters	many parameters	many parameters
□	□	□

A.4.5.3. How are model anomalies (see section A.5.5) discovered and discussed in the procedures for developing and testing the model?

Typically

incidental discovery	occasional attention	systematic routines
	to anomalies	to discover and discuss
□	□	□

A.4.5.4. Does one research group have an effective monopoly of access to the model? What are the mechanisms for scientific criticism (from peers in the science community)?

Check all that apply:

◊ The source code is public.

◊ Other groups use the model.

◊ The model is well documented in the literature.

◊ Specialized hardware and software are not required to run the model.

◊ There is active collaboration with outside groups in designing and analysing model runs.

◊ Other. Describe.

A.5. Interface with users

This section is intended to address interactions between model groups and those who use the model or its output. Issues covered include that of how well model output forms match the requirements of users, the management of model anomalies, and the levels of expertise required to use the model.

A.5.1. Scale

A.5.1.1. What is the models spatial resolution?

□

A.5.1.2. What is the models temporal resolution?

□

A.5.1.3. What is the models time horizon?

□

A.5.1.4. How do these scales relate to the needs of users of model output?

	too coarse	about right	finer than required
spatial resolution	□	□	□
temporal resolution	□	□	□
time horizon	□	□	□

A.5.2. Choice of output metrics

A.5.2.1. What indicators have been chosen to represent the outcome of model runs for this application?
List the main ones, with a brief note on their relevance to users.

A.5.2.2. Are these indicators the most appropriate metrics for users for this problem?
If not, list appropriate metrics and describe the relationships between what you do use and these metrics.

A.5.3. Tests for pseudo-precision

A.5.3.1. What level of accuracy for each metric is consistent with levels of uncertainty in the model?

metric	order of magnitude	a factor of 2	better than 10%
	□	□	□
	□	□	□
	□	□	□

A.5.3.2. Is this inherent accuracy reflected in the precision of numerical outputs?
If not, why not?

A.5.3.3. What is the relation of this accuracy to the requirements of users?

under-precise	a good match	over-precise
□	□	□

A.5.4. Tests for pseudo-imprecision

A.5.4.1 Have results been expressed so vaguely that they are immune from refutation or even criticism?
How would you characterize the relationship between the precision of model results and available data?

results are too vague to be refuted	results on the border of precision needed to allow refutation	results are precise enough to be refuted
□	□	□

A.5.5. Management of anomalies

A model anomaly is a model result that does not conform to the accepted standard of plausibility for model response. A model result may be anomalous relative to other models or to expectations from theory or observation. By this definition, anomalies are not necessarily errors. Anomalies seem implausible relative to the standards employed, but the standards may turn out to be wrong.

A.5.5.1. Describe some model anomalies from the current model.
These may be either current anomalies or those uncovered during the course of developing the model.

A.5.5.2. Who is included in the peer community for discussing model anomalies?
Check all that apply:

◊ Your immediate model group.
◊ Other groups at your institution.
◊ Other model groups in this field.
◊ The wider community in this field.
◊ User groups in other fields.
◊ The general public.
◊ Other. Describe.

A.5.5.3. In relation to user groups and the public, how are unresolved anomalies in the model or application managed?

	secrecy	tact	openness
users	□	□	□
public	□	□	□

A.5.6. Expertise

A.5.6.1. What levels of expertise and skill are required for competent use of the model by users?

minimal	moderate	considerable
□	□	□

A.5.6.2. What procedures are there for assessing the competence of those who use the model and its output?

minimal contact with users	moderate liason with users	close liason and follow through
□	□	□

A.6. Use of the models in policy

This section addresses a variety of issues in the presentation and use of model results in the policy process. This includes issues such as incentives related to results, how stakeholder perspectives have been addressed, and how much stakeholders understand of the basis of key model results.

A.6.1. Stakeholders

A.6.1.1. At what stage in the model experiment process were relevant stakeholders identified?

prior to running experiments	during the course of experiments	after running experiments
□	□	□

A.6.1.2. What expertise do stakeholders have on this issue?

minimal	moderate	substantial
□	□	□

Describe.

A.6.1.3. What was the level of stakeholder participation in the problem formulation phase (model experiment design)?

minimal	moderate	substantial
□	□	□

A.6.1.4. Have rival problem formulations been considered?
Briefly describe them and their implications for this issue.

A.6.2. Results

A.6.2.1. What is the level of accuracy required for model results to be useful in the policy process?

order of magnitude	a factor of 2	better than 10%
□	□	□

A.6.2.2. How do the requirements for accuracy in the policy process compare with the accuracy achieved by the model (indicated in question A.5.3.1)?

model results too coarse for this application	about the required level of accuracy	model more than accurate enough
□	□	□

A.6.2.3. Does the model give useful answers to the problem posed?

not relevant or plausible	relevant but with unknown plausibility	relevant and plausible	provides relevant and compelling results
□	□	□	□

A.6.2.4. How are the model results used in the policy process?
Check all that apply:

◊ Substantively, influencing contents of a policy proposal or implementation.

◊ Rhetorically, pro or con a policy.

◊ Primarily for community-building among modellers or users.

◊ Other. Describe.

If this assessment differs substantially from your response on how models *ought* to be used in the policy process (question A.3.4.5), what are the main reasons?

A.6.2.5. Are there investments in particular model results by modellers and/or users and stakeholders?
Identify and describe.

Modellers:
Users:
Stakeholders:

A.6.2.6. Is there evidence or suspicion of WYGIWYN – ‘What You Get Is What You Need’, as a policy of modellers in relation to funders or users and stakeholders?

What is the relationship between results and the interests of the following groups on this issue. For each group below, answer for the major entity in the group or for a typical entity.

	results typically at odds	some results convergent	results typically convergent
funders	□	□	□
users	□	□	□
stakeholders	□	□	□

A.6.2.7. Is probabilistic (or other) information used in communicating uncertainty about results?

minimal uncertainty information given	qualitative description of uncertainty ranges	error bars estimated or range given	results given as pdf’s
□	□	□	□

A.6.3. Transparency in the policy process

A.6.3.1. Has the model been designed to enable scrutiny and testing by (or on behalf of) all stakeholders in a policy debate?

minimal use or inspection of model by stakeholders	moderate use of model by stakeholders	stakeholders frequently exercise the model
□	□	□

A.6.3.2. Are some potential stakeholders excluded by the requirements of the model for bases of knowledge, expertise, software, and hardware?

model too complex and/or hardware too specialized for outside use	some have sufficient resources to exercise the model	model simple enough and portable enough to allow virtual open access
□	□	□

A.6.3.3. Have relevant value judgements in the model been identified and made explicit in presenting results?

occasionally articulated in presentations	often articulated in presentations	clearly identified in most public presentations
□	□	□

A.6.3.4. Is it clear to users what the effects of the different value choices are?

mostly unaware of implications of different choices	partially aware of implications	can describe most value implications with reference to model formulations
□	□	□

A.6.3.5. Can alternative value choices be implemented and evaluated with the model at a user’s request?

rarely	depending on the details of value formulation	readily
□	□	□

A.6.4. Other

6.4.1. Are there any other relevant properties of the model that have not been covered in this checkist?

A.7. Summary assessment

This section covers both a holistic assessment of the use of the model and provides for summaries of the results of previous sections. The results summary is given in the form of a list of potential pitfalls. The pitfalls describe issues that may affect the maintenance of quality in using the model for the intended purpose.

A.7.1. Overall assessment

A.7.1.1. For this particular problem, model results can be used:
Provide your subjective overall assessment from the list below.

○ With High Confidence
○ With Confidence
○ With Caution
○ With Extreme Caution

What were the most important factors that led you to choose this ranking?

A.7.2. Potential pitfalls

The responses to checklist questions can be used to generate a set of potential pitfalls as a diagnostic aid. The pitfalls indicate areas where quality is at stake and where there are more likely to be problems encountered in applying the model. The version of the model checklist running on the web (http://www.nusap.net) generates a list of pitfalls automatically on the basis of checklist responses.

A.7.3. Caveat utor

Checklists such as this are an exercise in quality control. That always raises the issue of who will quality control the quality controllers? For many complex model domains quality *control* is an elusive goal, since quality cannot be completely tamed and managed. Thus, we prefer to view a checklist such as this as an exercise in quality *assistance,* which is necessarily limited and cursory. One should always seek additional means to assist in the quality process.

References

[1] W. Keepin, Review of global energy and carbon dioxide projections, Ann. Rev. Energy 3(11) (1986) 357–392.

[2] S. Funtowicz and J. Ravetz, *Uncertainty and Quality in Science for Policy* (Kluwer, Dordrecht, 1990) 229 pp.

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

[4] N. Oreskes, K. Shrader-Frechette and K. Belitz, Verification, validation, and confirmation of numerical models in the earth sciences, Science 263(3) (1994) 641–646.

[5] J. Risbey, M. Kandlikar and A. Patwardhan, Assessing integrated assessments, Clim. Change 34(3–4) (1996) 369–395.

[6] B. Beck, Model evaluation and performance, in: *Encyclopedia of Environmetrics*, Vol. 3 (John Wiley & Sons, New York, 2002) pp. 1275–1279.

[7] 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) 26 pp.

[8] C. Mann, Why software is so bad, Tech. Review 1(4) (2002) 33–38.

[9] J. van der Sluijs, J. Risbey and J. Ravetz, Uncertainty assessment of VOC emissions from paint in the Netherlands using the nusap system, Env. Mod. Ass. (2004) in press.

[10] J. Ravetz, *Scientific Knowledge and Its Social Problems* (Clarendon Press, Oxford, 1971), Reprint: Transaction, New Brunswick NJ (1996) 449 pp.

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

[12] N. Nakićenović, J. Alcamo, G. Davis, B. de Vries et al., Special report on emissions scenarios: A special report of the Intergovernmental Panel on Climate Change, Cambridge Univ. Press, Cambridge, UK (2000) 599 pp.

[13] B. De Vries, 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) (2000) 137–174.

[14] B. De Vries, 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.

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

[16] J. Alcamo, R. Leemans and E. Kreileman, eds., *Global Change Scenarios for the 21st Century. Results from the IMAGE 2.1 Model* (Elsevier Science, London, 1998) 572 pp.

[17] S. Funtowicz and J. Ravetz, The worth of a songbird: ecological economics as a post-normal science, Ecol. Econ. 3(10) (1994) 197–207.

[18] S. Schneider, Integrated assessment modeling of global climate change: transparent rational tool for policy making or opaque screen hiding value-laden assumptions, Env. Modeling and Assessment 2(6) (1997) 229–249.

[19] P. Kloprogge and J. van der Sluijs, Choice processes in modelling for policy support, in: *Proceedings of the International Environmental Modelling and Software Society*, Vol. 1, Lugano, June 2002, IEMSS, pp. 96–101.

[20] J. van der Sluijs, J. Risbey, S. Corral Quintana and J. Ravetz, Uncertainty management in complex models: the NUSAP method, in: *Proceedings of the International Environmental Modelling and Software Society*, Vol. 2, Lugano, June 2002, IEMSS, pp. 13–18.

[21] United Nations, United Nations Framework Convention on Climate Change. Text available at http://unfccc.int (1992).

[22] A. Petersen, P. Janssen, J. van der Sluijs, J. Risbey and J. Ravetz, RIVM/MNP guidance for uncertainty assessment and communication: Mini-checklist and quickscan questionnaire, Technical report, Netherlands Environmental Assessment Agency (2003) 15 pp.