A Checklist for Quality Assistance in Environmental Modelling

James Risbey, Jeroen van der Sluijs, Jerry Ravetz, and Peter Janssen

NW&S-E-2001-11
ISBN 90-73958-65-2
July 2001

Acknowledgments

This research was undertaken as part of the project 'Uncertainty Assessment IMAGE 2', supported by the Dutch National Research Program on Global Air Pollution and Climate Change (contract no. 954267). Further support was provided by a visitor grant from NWO (B 76-220). Substantial contributions to this research were made by a number of people, including: Silvio Funtowicz, Serafin Corral, Pita Verwey, Penny Kloprogge, Jose Potting, Jos Dekker, Heleen Groeneberg, Arthur Petersen, and Milind Kandlikar.

A Checklist for Quality Assistance in Environmental Modelling

James Risbey[1], Jeroen van der Sluijs[1], Jerry Ravetz[2], and Peter Janssen[3]
Department of Science, Technology, and Society
Utrecht University, Utrecht 2001
Report No. NW&S-E-2001-11
ISBN 90-73958-65-2

[1]Department of Science, Technology, and Society
Utrecht University
Padualaan 14
3584 CH Utrecht
The Netherlands
Phone: +31 30 253-7600
Fax: +31 30 253-7601
http://www.chem.uu.nl/nws/nws.html

[2]Research Methods Consultancy
7th floor, Methodist Church House
27 Marylebone Road, London NW1 5JS UK

[3]RIVM
P.O. Box 1, 3720 BA Bilthoven
The Netherlands

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. The first section of the checklist proper (section 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 4 addresses 'internal' quality issues, which refers to the processes for developing, testing, and running the model practiced within the modelling group. Section 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 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.

2 Screening questions

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:

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
☐	☐

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
☐	☐

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

peers ☐
users ☐
stakeholders ☐

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
☐	☐	☐

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

community widely discredited	mixed acceptance	community widely accepted
☐	☐	☐

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
☐	☐	☐

2.2 Which parts of the checklist are potentially useful?

Section 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.

2.2.1 Internal Strength

Section 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

2.2.2 Interface with users

Section 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

2.2.3 Use in policy

Section 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

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.

3.1 Model name:

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

3.2 Intended function or application

3.2.1 Describe the problem being addressed

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

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.

3.3 Intended users

Identify the users of model results and interested stakeholders.

3.4 Problem domain

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
	☐	☐	☐
	☐	☐	☐
	☐	☐	☐
	☐	☐	☐

3.4.2 List any pertinent facts that are in dispute?

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.

3.4.4 Who funds your groups research on this issue?

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.

4 Assessment of Internal Strength

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

4.1 Parametric uncertainty and sensitivity

4.1.1 Has the strength of the input data been assessed?

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

4.1.2 Have the key parameters (governing spread in model output) been identified?

List them.

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 ☐

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 ☐

4.2 Structural uncertainty assessment

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

Describe them.

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

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

trivially ☐ moderately ☐ radically ☐

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 ☐

4.2.5 How was the system boundary defined?

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

4.2.6 Have the consequences of alternative boundary choices been examined?

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

trivial ☐ moderate ☐ radical ☐

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

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
	☐	☐	☐
	☐	☐	☐
	☐	☐	☐

4.3 Validation

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.

4.4 Robustness

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

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 assumptions	some of the major assumptions	most of the major assumptions
☐	☐	☐

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%
☐	☐	☐

4.5 Model Development Practices

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.

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 on a few parameters	occasionally, including many parameters	often, including many parameters
☐	☐	☐

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

Typically

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

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.

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.

5.1 Scale

5.1.1 What is the models spatial resolution?

5.1.2 What is the models temporal resolution?

5.1.3 What is the models time horizon?

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	☐	☐	☐

5.2 Choice of output metrics

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.

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.

5.3 Tests for pseudo-precision

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%
	☐	☐	☐
	☐	☐	☐
	☐	☐	☐

5.3.2 Is this inherent accuracy reflected in the precision of numerical outputs?

If not, why not?

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

under-precise ▭ a good match ▭ over-precise ▭

5.4 Tests for pseudo-imprecision

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 ▭

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.

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.

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.

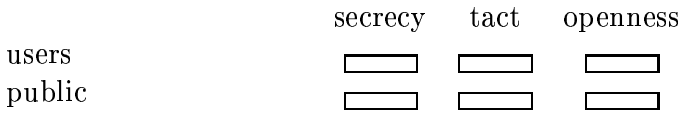
♢ The wider community in this field.

♢ User groups in other fields.

♢ The general public.

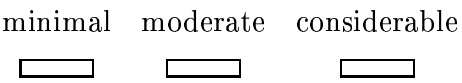
♢ Other. Describe.

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

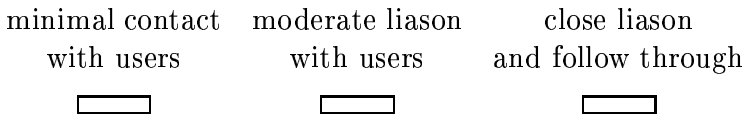


5.6 Expertise

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



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



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.

6.1 Stakeholders

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 ☐

6.1.2 What expertise do stakeholders have on this issue?

minimal ☐ moderate ☐ substantial ☐

Describe.

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

minimal ☐ moderate ☐ substantial ☐

6.1.4 Have rival problem formulations been considered?

Briefly describe them and their implications for this issue.

6.2 Results

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% ☐

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

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

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
☐	☐	☐	☐

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 3.4.5), what are the main reasons?

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

Identify and describe.

Modellers:

Users:

Stakeholders:

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	☐	☐	☐

6.2.7 Is probabilitic (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
☐	☐	☐	☐

6.3 Transparency in the policy process

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
☐	☐	☐

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
☐	☐	☐

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
☐	☐	☐

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
☐	☐	☐

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
☐	☐	☐

6.4 Other

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

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 maintainance of quality in using the model for the intended purpose.

7.1 Overall assessment

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?

7.2 Potential Pitfalls

Some of the potential pitfalls identified from your responses will be listed here.

7.3 Caveat

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 can not 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.