Reframing the Reliability of Models
Moving From Error to Quality for Use

Arthur C. Petersen[1,2] and Leonard A. Smith[2]

[1] Netherlands Environmental Assessment Agency (PBL)
[2] Centre for the Analysis of Time Series (CATS), London School of Economics and Political Science (LSE)

Draft Paper for ESF workshop
Exploring Epistemic Shifts in Computer Based Environmental Sciences
June 10–12th , Aarhus University, Denmark
(Contribution to session 3, 'Standards of good science')

1. Introduction

Scientific computer simulation—here defined as involving a mathematical model that is implemented on a computer and that imitates real-world processes—is portrayed by some philosophers of science as a new method of doing science, besides theorizing and experimentation (e.g., Rohrlich 1991; Humphreys 1994; Keller 2003). Science studies generally seems to support this conclusion from an historical or sociological perspective (e.g., Galison 1996; Dowling 1999). Two major reasons are typically given for why simulation should be considered qualitatively different. First, it is claimed that simulations make it possible to 'experiment' with theories in a new way (e.g., Dowling 1999: 271). Second, simulation enables us to extend our limited mathematical abilities so that we can now perform calculations that were hitherto unfeasible. Thus, we can both construct new theories using computer simulation and calculate the consequences of old theories.[1] An example of the former category is the application of cellular automata in biology (Rohrlich 1991; Keller 2003). And the latter category is exemplified by the forecasting of the weather on the basis of the well-established nonlinear equations of fluid dynamics, the Navier–Stokes equations.

Although there is no fundamental difference between the computability of problems before and after computers became available (an argument against putting simulation in a philosophically distinct category), through the introduction of the computer an actual barrier in scientific practice to the large-scale use of numerical mathematics—that is, the limited speed with which humans, even if aided by mechanical machines, can do calculations—was removed. We agree with Paul Humphreys:

> While much of philosophy of science is concerned with what can be done in principle, for the issue of scientific progress what is important is what can be done in practice at any given stage of scientific development (Humphreys 1991: 499).

Science studies—which includes at least philosophical, historical, sociological and anthropological approaches—can shed light on all kinds of changes in scientific practice

[1] Both 'new' aspects of scientific research are made possible by the high speed with which computers perform calculations. And both aspects are interrelated: we can derive new theories by 'playing' with variants of old theories and the calculation of results for the new theories is often unfeasible without using the computer.

that have resulted from the explosive growth of computer-based science. Here we argue that in particular the question of how the 'reliability' of models is conceived and established merits further interdisciplinary study, not only involving the science studies community but also involving practitioners. We consider the environmental sciences—and in particular climate modelling—as a prime location where modelling practices need to be reflected upon, both for theoretical and practical reasons. In this paper we address the question of how the 'reliability' of models could be framed in different ways.

	Sorts of uncertainty						
UNCERTAINTY MATRIX	**Nature of uncertainty**		**Range of uncertainty (inexactness/imprecision or unreliability$_1$/inaccuracy)**		**Recognised ignorance**	**Methodological unreliability (unreliability$_2$)** • Theoretical basis • Empirical basis • Comparison with other simulations • Peer consensus	**Value diversity** • General epistemic • Discipline-bound epistemic • Socio-cultural • Practical
Location/source of uncertainty ↓	**Epistemic uncertainty**	**Ontic uncertainty / indeterminacy**	**Statistical uncertainty (range+ chance)**	**Scenario uncertainty (range of 'what-if' options)**			
Conceptual model							
Mathematical model — **Model structure**							
Mathematical model — **Model parameters**							
Model inputs (input data, input scenarios)							
Technical model implementation (software and hardware implementation)							
Processed output data and their interpretation							

Figure 1. Typology of uncertainty in computer simulation (Petersen 2006)

2. Establishing Reliability Through Error Analysis

A scientific claim based on results from computer simulation may express a range of uncertainty. This range may in turn derive from uncertainty sources at different locations (see Fig. 1). In science, uncertainty ranges come in two types: statistical uncertainty and scenario uncertainty ranges. A 'statistical uncertainty' range can be given for the uncertainties which can be adequately expressed in statistical terms, e.g., as a range with associated probability (for example, uncertainties in model-parameter estimates). In the natural sciences, scientists generally refer to this category of uncertainty, thereby often implicitly assuming that the model relations involved offer adequate descriptions of the real system under study, and that the (calibration-)data employed are representative of the situation under study. Two different statistical paradigms are available for characterising probabilities: the frequentist and the Bayesian paradigm. In frequentist statistics, probabilities are considered to be 'objective' and are based on the empirically determined frequency of occurrence of events. In Bayesian statistics, probabilities are 'subjective', based on expert judgement and reflecting all information that is available on a particular event (not necessarily the frequency of occurrence).[2]

[2] Although the IPCC is not very explicit about it and also no rigorous Bayesian updating procedure has been followed, its conclusion that most of the observed warming over the last 50 years is attributable to the anthropogenic emissions of greenhouse gases was qualified in a Bayesian framework: in 2001 the IPCC

However, 'deeper' forms of uncertainty are often at play. These cannot be expressed statistically but can sometimes be expressed by a range. Such a range is then called a 'scenario uncertainty' range. Scenario uncertainties cannot be adequately depicted in terms of chances or probabilities, but can only be specified in terms of (a range of) equally plausible events. For such uncertainties to specify a degree of probability or belief is meaningless, since the mechanisms which lead to the events are not sufficiently known. Scenario uncertainties are often construed in terms of 'what-if' statements. In principle, it is possible that uncertainties which are first expressed as scenario uncertainties later switch to the category of statistical uncertainty if more becomes known about the system processes.[3] There are claims that this has happened with simulation-based estimates of the sensitivity of climate to greenhouse-gas increases, the 'climate sensitivity'.[4] In recent years, more and more articles have appeared in the literature that include statistical characterizations, through probability density functions, of the climate sensitivity (e.g, Murphy et al. 2004). However, some refuse to label these model-based distributions as if they were probabilities of climate sensitivity (Stainforth et al. 2005). The IPCC took a middle road, labeling the y-axis of such graphs 'Relative Probability'.

The measure of the spread of both types of uncertainty range is either inexactness,[5] also called 'imprecision', which gradually moves from exact (small range) to inexact (large range), or unreliability$_1$ (defined below), also called 'inaccuracy', which gradually moves from accurate (small range) to inaccurate (large range).[6] Ranges of uncertainty can derive from any source of uncertainty in scientific practice, including model structure. Thus, in principle, 'systematic error' can be represented in the typology as statistical uncertainty arising from the model structure.

The statistical uncertainty range can be qualified by information about the statistical reliability (reliability$_1$). The term '(un)reliability' is frequently used in science without explicit mention of what precisely is meant. We argue here that reliability must be understood as *reliability for a purpose*. Following Petersen (2006), two different notions of 'reliability' are distinguished, denoted by reliability$_1$ (statistically characterized reliability) and reliability$_2$ (methodologically characterized reliability). As Parker (2009) has shown, in order to establish the 'adequacy-for-purpose' of a model, scientists rely not simply on the statistical promixity of model results to, for instance, an historical dataset of the quantity of interest (since in that way the reliability$_1$ for predicting the future cannot be established), but use a much more elaborate argumentation. This argumentation includes, we claim, an assessment of the reliability$_2$ of the model, for instance of the

---

qualified this conclusion as 'likely' (defined as a 'judgmental estimate' of a 66–90% chance of being correct) and in 2007 as 'very likely' (a >90% chance).

[3] The switch can work the other way round (from statistical uncertainty to scenario uncertainty), if one later realizes that too little is known.

[4] The 'climate sensitivity' is defined as the equilibrium global surface temperature increase for a doubling of the CO_2 concentration.

[5] The term 'exactness' (or 'inexactness') has many meanings (Kirschenmann 1982). I propose to use this one in accordance with Funtowicz and Ravetz (1990).

[6] While Funtowicz and Ravetz considered a 'spread' to derive typically from 'random error' (e.g., Funtowicz and Ravetz 1990: 23, 70), the interpretation of 'range' in the typology used here is much broader and encompasses all types of statistical and scenario-uncertainty ranges.

methodological quality of the representation of a particular dynamic process that is thought to be of importance for its use, e.g., for modelling particular future changes.

Let us explore both notions of reliability a bit more, starting with reliability$_1$. For scientific simulation laboratory practice, 'reliability$_1$' is defined as follows: the 'reliability$_1$' of a simulation is the extent to which the simulation yields accurate results in a given domain. It is important here to distinguish between 'accuracy' and 'precision' (see Hon 1989: 474). Accuracy refers to the closeness of the simulation result to the 'true' value of the sought physical quantity, whereas precision indicates the closeness with which the simulation results agree with one another, independently of their relations to the 'true' value. 'Accuracy thus implies precision but the converse is not necessarily true' (Hon 1989: 474).[7] Traditionally, the distinction between 'systematic' and 'random' error is taken to correspond with the distinction between 'accuracy' and 'precision' (Hon 1989: 474). Since systematic and random error are both statistical notions, Petersen (2006: 55) proposed to dissociate these two dichotomies from each other, so that all sources of error may be assessed in terms of their impact on the accuracy and the precision of the results.[8]

There is an epistemological and practical problem with maintaining a strong focus on the statistical reliability of models, however. We know that models are not perfect and never will be perfect. Especially when extrapolating models into the unknown, we wish 'both to use the most reliable model available and to have an idea of how reliable that model is' (Smith 2002: 2491), but the statistical reliability cannot be established. There is no statistical fix here; also, we should not confuse range of outcomes of a diversity of models in ensemble projections, such as used by the IPCC, with a statistical measure of uncertainty. This does not mean that the information in models cannot be used, but it does imply that the 'reliability' of models needs to be assessed in the context of their use.

3. Establishing Reliability Through Assessment of Quality for Use

Models, when they do apply, will hold only in certain circumstances. We may, however, be able to identify shortcomings of our model even within the known circumstances and thereby increase our understanding (Smith 2002). As was observed in the previous section, a major limitation of the statistical definition of reliability is that it is often not possible to establish the accuracy of the results of a simulation or to quantitatively assess

[7] For probability density functions as results of simulation, which obviously can be accurate but need not be sharp, this statement should be interpreted to mean: 'An accurate PDF thus implies a precise PDF [that is, PDFs lie close to each other] but the converse is not necessarily true'.

[8] The theory of statistical error analysis, as is developed for instance by Deborah Mayo (1996) for experimentation, is too limited for a proper analysis of uncertainty in science. Both in the experimental and the simulation laboratory it is often not a straightforward exercise to determine the reliability$_1$ of a simulation: there are many different factors involved that could cause error. The standard view of error in scientific textbooks holds that, 'apart from random errors, all experimental errors can be eliminated'—a view which is grossly mistaken, however, since the complexity of actual reality typically prevents systematic errors from being reducible to zero (Hon 1989: 476). The distinction between systematic and random errors is only mathematically based and has only limited value for actually determining error in scientific practice. We therefore agree with Hon (2003: 191-193) that the theory behind the concepts of random and systematic errors is purely statistical and not related to the locations and other dimensions of uncertainty.

the impacts of different sources of uncertainty. Furthermore, disagreement (in distribution) between different modelling strategies would ague against the reliability of some, if not all, of them. 'Reliability' then will have to be defined in more pragmatic terms. In those cases, one may instead have recourse to *qualitative* judgments of the relevant procedures, the methodological quality given the purpose of use. A methodological definition of reliability, denoted by reliability$_2$, can be given as follows: the 'reliability$_2$' of a simulation is the extent to which the simulation has methodological quality. The methodological quality of a simulation derives from the methodological quality of the different elements in simulation practice, *given the purpose of the model*. The methodological quality of a simulation, for example, depends not only on how adequately the theoretical understanding of the phenomena of interest is reflected in the model structure, but also, for instance, on the quality of the initial and boundary conditions used as input data to the model; the numerical algorithms; the procedures used for implementing the model in software; the statistical analysis of the output data; etc.

While the range of uncertainty is a quantitative dimension of uncertainty, the other five dimensions (including location) of uncertainty (see Fig. 1) are qualitative. Since methodological quality is a qualitative dimension and the (variable) judgment and best practice of the scientific community provides a reference, determining the methodological quality of a claim is not usually a straightforward affair either. It depends, for instance, on how broadly one construes the relevant scientific community and what one perceives as the purpose of the model. The broader the community, the more likely it is that the different epistemic values held by different groups of experts could influence the assessment of methodological quality. Criteria such as (1) theoretical basis, (2) empirical basis, (3) comparison with other simulations and (4) acceptance/support within and outside the direct peer community can be used for assessing and expressing the level of reliability$_2$ (see Petersen 2006: 58–62).

4. Conclusion

Given the presence of many different ways the reliability of models is established in scientific practices and the importance attached to the assessed reliability in particular decision-making context, such as is in climate-policy making, it is important for science studies to further investigate the notion of reliability. The present paper has just presented some first analytical steps. Subsequent research should preferably be done in a multi-disciplinary fashion, thus combining philosophical, sociological, anthropological and historical expertise.

References

Dowling, D. (1999), 'Experimenting on theories', *Science in Context* 12: 261–273.

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

Galison, P. (1997), *Image and Logic: A Material Culture of Microphysics*, Chicago: Chicago University Press.

Hon, G. (1989), 'Towards a typology of experimental errors: an epistemological view', *Studies in History and Philosophy of Science* 20: 469–504.

Hon, G. (2003), 'The idols of experiment: transcending the 'etc. list'', in H. Radder (ed.), *The Philosophy of Scientific Experimentation*, Pittsburgh: Pittsburgh University Press, pp. 174–197.

Humphreys, P. (1991), 'Computer simulations', in A. Fine, M. Forbes and L. Wessels (eds.), *PSA 1990, Vol. 2: Symposium and Invited Papers*, East Lansing: Philosophy of Science Association, pp. 497–506.

Humphreys, P. (1994), 'Numerical experimentation', in P. Humphreys (ed.), *Patrick Suppes: Scientific Philosopher, Vol. 2*, Dordrecht: Kluwer Academic Publishers, pp. 103–121.

Keller, E.F. (2003), 'Models, simulation, and 'computer experiments'', in H. Radder (ed.), *The Philosophy of Scientific Experimentation*, Pittsburgh: Pittsburgh University Press, pp. 198–215.

Kirschenmann, P.P. (1982), 'Some thoughts on the ideal of exactness in science and philosophy', in J. Agassi and R.S. Cohen (eds.), *Scientific Philosophy Today: Essays in Honor of Mario Bunge*, Dordrecht: D. Reidel Publishing Company, 85–98.

Murphy, J.M. et al. (2004), 'Quantifying uncertainties in climate change from a large ensemble of general circulation model predictions', *Nature* 430: 768–772.

Parker, W.S. (2009), 'Confirmation and adequacy-for-purpose in climate modelling', *Proceedings of the Aristotelian Society Supplementary Volume* 83: 233–249.

Petersen, A.C. (2006), *Simulating Nature: A Philosophical Study of Computer-Simulation Uncertainties and Their Role in Climate Science and Policy Advice*. Het Spinhuis Publishers, Apeldoorn and Antwerp. Download from http://hdl.handle.net/1871/11385.

Rohrlich, F. (1991), 'Computer simulation in the physical sciences', in A. Fine, M. Forbes and L. Wessels (eds.), *PSA 1990, Vol. 2: Symposium and Invited Papers*, East Lansing: Philosophy of Science Association, pp. 497–506.

Smith, L.A. (2002), 'What might we learn from climate forecasts?', *Proceedings of the National Academy of Sciences of the United States of America* 99 (Suppl. 1): 2487–2492.

Stainforth, D.A., Aina, T., Christensen, C., Collins, M., Frame, D.J., Kettleborough, J.A., Knight, S., Martin, A., Murphy, J., Piani, C., Sexton, D., Smith, L., Spicer, R.A., Thorpe, A.J., Webb, M.J. and Allen, M.R. (2005), 'Evaluating uncertainty in the climate response to changing levels of greenhouse gases', *Nature* 433 (7024): 403–406.