Synthese (2008) 163:371–384
DOI 10.1007/s11229-007-9296-0

Computer simulation through an error-statistical lens

Wendy S. Parker

Received: 9 November 2007 / Accepted: 9 November 2007 / Published online: 20 February 2008

Abstract After showing how Deborah Mayo's error-statistical philosophy of science might be applied to address important questions about the evidential status of computer simulation results, I argue that an error-statistical perspective offers an interesting new way of thinking about computer simulation models and has the potential to significantly improve the practice of simulation model evaluation. Though intended primarily as a contribution to the epistemology of simulation, the analysis also serves to fill in details of Mayo's epistemology of experiment.

Keywords Computer simulation · Evidence · Error statistics · Climate models

1 Introduction

Computer simulation models have emerged as an important research tool in many scientific fields. Painting with a broad brush, we can identify at least two epistemic functions that computer simulation models might serve. First, they might serve as heuristic tools: interaction with computer simulation models might help scientists to arrive at novel hypotheses to be subjected to further investigation via observation and experiment. There seems to be broad agreement among scientists and philosophers alike that computer simulation models can have this heuristic value.

Second, computer simulation models might serve as evidential resources: they might be used in investigations that are meant to provide good evidence for hypotheses about real-world target systems. Surveying actual modeling practice, it seems clear that computer simulation models sometimes are developed with this goal in mind. Obvious examples, which will be revisited below, include the development of

W. S. Parker (✉)
Department of Philosophy, Ohio University, Ellis Hall 202, Athens, OH 45701, USA
e-mail: parkerw@ohio.edu

computer models of earth's atmosphere and climate system, which scientists hope will be able to provide accurate information regarding tomorrow's weather and the next century's climate (respectively). But can weather and climate models, or computer simulation models in other fields, really deliver good evidence regarding hypotheses about real-world target systems? If so, when? When it comes to questions like these, little or no consensus has been reached; indeed, there is a widely-recognized need for more discussion and analysis (e.g. Oreskes et al. 1994; Rykiel 1996; Winsberg 1999a; Beck 2002).

The present paper aims to contribute to the discussion by grappling with two fundamental questions about the evidential status of computer simulation results:

(a) What is required for computer simulation results to constitute good evidence for hypotheses about real-world target systems?
(b) How do we go about determining whether the results of some simulation study constitute such evidence in any particular case?

The approach taken involves viewing computer simulation modeling through the lens of a particular account of scientific evidence, namely, the error-statistical account recently developed by Deborah Mayo (1996, 2000, 2005). After drawing on this account to provide preliminary answers to (a) and (b), I explain why adopting an error-statistical perspective could be particularly valuable when it comes to understanding and using computer simulation models as evidential resources.

Section 2 introduces the error-statistical framework and its key notion of severe testing and explains why Mayo's discussion itself requires an analysis of the evidential status of computer simulation models. In Sect. 3, I apply the error-statistical framework to computer simulation modeling to arrive at preliminary responses to (a) and (b) and then offer some remarks on the prospects for using computer simulation results as evidence for hypotheses about real-world target systems. Throughout, key points are illustrated in the context of weather and climate modeling. Despite remaining questions about the use of simulation results as evidence, Sect. 4 argues that adopting an error-statistical perspective on computer simulation modeling would be valuable for at least two reasons: it would offer an interesting new way of thinking about computer simulation studies and it would promote important and healthy changes in the practice of simulation model evaluation. Section 5 offers some concluding remarks.

2 Mayo's error-statistical philosophy of science

Mayo's (1996, 2000, 2005) account of scientific evidence is built around the notion of a severe test. A *severe test* of some hypothesis H is a procedure that has a high probability of rejecting H, if and only if H is false. We say that H *passes a severe test* with results e just in case: (i) e fit H, for some suitable notion of fit; and (ii) it is very unlikely that the test procedure would produce e that fit so well with H, if H is false. If H does pass a severe test with results e, then e are *good evidence for*—or a *good indication of*—H. Put more informally: we have good evidence for H just in case a procedure that almost surely would have indicated H to be in error, were H actually erroneous, nevertheless does not indicate that H is in error (For more details on these definitions and requirements, see Mayo 1996, 2000, 2005).

In conjunction with her account of evidence, Mayo has offered an extended analysis of how traditional experimental inquiry can provide good evidence for scientific hypotheses. In a nutshell, it is because scientists have become shrewd inquisitors of experimental error—they have identified canonical sources of error that can impact experimental results and have developed a variety of tests that probe for the presence of those sources of error. Such sources of error include, but are not limited to: a failure to meet the design assumptions of the experiment, a malfunctioning experimental apparatus, a biased data processing technique, and a failure to adequately control for confounding factors (see Mayo 1996, Chap. 5 for more details). On Mayo's view, we must be able to argue that such sources of experimental error were absent from our experiment and thus have not impacted our results (or, if some of them were present, that they have not impacted the results by more than a specified amount), before we can claim that our results constitute good evidence regarding some primary hypothesis of interest (e.g., a hypothesis about the efficacy of a drug in a clinical trial). To be in a position to make such an argument, we typically must conduct a battery of lower-level severe tests, each of which is designed to probe for the presence of one or more specific sources of error, such as an instrument malfunction or the presence of a confounding factor.

The statistical part of Mayo's error-statistical approach is tied to the details of severe testing, and it is grounded in a frequentist interpretation of probability. She argues that formal statistical tools and concepts are especially useful when it comes to probing for error, not least because they can help us to determine what we would be more and less likely to observe when carrying out some test procedure, if a particular source of error were present (1996, p. 164). With this information, we may be in a position to draw conclusions about the presence and/or impacts of that source of error. To take a very simple example, if we know that 99% of the time the temperature registered by the thermometer used in our experiment is within 1° F of the true temperature, then we can estimate how likely it is that random measurement error would lead us to erroneously accept, on the basis of the particular experimental results we obtained, a particular hypothesis that we set out to test concerning the temperature of a substance. As the example illustrates, statistical tools and concepts don't give us answers "from thin air" (Mayo 1996, p. 96)—we have to draw on our knowledge of the subject-matter at hand as well—but in conjunction with that knowledge, statistical concepts and tools can play a valuable role in helping us to design severe tests and to decide whether they have been passed (see also Mayo 1996, pp. 449–462). It is important to recognize, however, that although formal statistical analysis has particular value on Mayo's view, she does not claim that it is essential in severe testing for error. Instead, sometimes more qualitative arguments about what it would (and wouldn't) be like if a particular source of error were present in an experiment can be perfectly appropriate—what Mayo refers to as "informal" arguments from error (see e.g., Mayo 1996, pp. 12–13 and p. 138).

It is in discussing the need to model error in traditional experimental contexts that Mayo makes reference to computer simulation studies. In order to argue that some source of error was almost surely absent from an experiment, scientists sometimes carry out computer simulations to help them estimate what they would be more and less likely to observe if that source of error were present in the experiment. Mayo illustrates

with the example of the discovery of neutral currents. Repeated observation of neutrino events without muons fit well with the hypothesis that neutral currents exist, but there was also the possibility that muons were actually present and simply not being detected due to inadequacies in the experimental apparatus. Computer simulations were carried out to estimate how many neutrino events without muons would be more and less likely to be observed if there were particular inadequacies in the muon detection apparatus, and the results of these simulations were used to reject the hypothesis that muons were present but simply escaping detection, on the grounds that the actual experimental results fit so poorly with what would be expected if the muons were really just escaping (see Mayo 1996, pp. 92–99 and pp. 162–164). Mayo's acceptance of the use of simulations in this way seems to commit her to the view that computer simulations sometimes do provide good evidence for hypotheses about the real-world systems they are chosen to represent. For, presumably, simulation results could be used in the way Mayo suggests—i.e. to argue that some source of error was absent from an experiment—only if they constituted good evidence for a hypothesis about what it would be like if that source of error were present in the experiment, and such a hypothesis *is* a hypothesis about a real-world target system, namely, the particular system that constitutes the experimental set-up.

So it seems that some understanding of how computer simulation results can constitute evidence for hypotheses about real-world target systems is needed even in the context of Mayo's analysis of traditional experimentation, because of the role that computer simulation studies sometime play in providing information about what it would be like if particular sources of error were present in a given experiment. As indicated previously, however, not all computer simulation studies are undertaken with the goal of aiding arguments in traditional experimental contexts; sometimes computer simulation studies are carried out on their own (unconnected with any traditional experiment) with the aim of providing evidence regarding natural systems outside of the laboratory, e.g., when computer simulation models are used to forecast tomorrow's weather. In both kinds of situation, it is hoped that computer simulation studies will provide good evidence regarding hypotheses about real-world target systems. An analysis of when computer simulation results constitute such evidence is needed.

3 Computer simulation through an error-statistical lens

Turning first to question (a): What is required for computer simulation results to constitute good evidence for hypotheses about real-world target systems? Viewing computer simulation studies as putative test procedures, Mayo's account of evidence can be applied directly to deliver the following answer. Simulation results constitute good evidence for some real-world hypothesis H just to the extent that:

(i) the results fit H; and
(ii) it is unlikely that the simulation study would deliver results that fit so well with H, if H is false.

Importantly, this implies that the procedure constituted by the simulation study must have a high probability of indicating that H is false, if H is in fact false; if the

procedure that constitutes a simulation study has little chance of revealing H to be in error even if it is in fact in error, then the simulation study cannot provide good evidence regarding H.

To illustrate, we can consider an example involving computer simulation models of earth's climate system. Suppose that these climate models project that at least moderate global warming will occur by the middle of the next century; does this count as good evidence for the hypothesis that at least moderate global warming will occur? According to the view just presented, the projections count as good evidence only if it is highly unlikely that climate models would deliver results that fit so well with the moderate-warming hypothesis, were it a false hypothesis about the next century's climate.

Reflecting on the climate modeling example leads us rather directly to (b): How do we go about determining whether the results of some simulation study constitute good evidence for a particular hypothesis? That is, how do we go about determining whether the simulation study and its results are such that, if H is false, then the simulation study almost surely would have delivered results that fit less well with H than the actual simulation results do? The strategy that Mayo presents for the case of traditional experimentation suggests one possibility, namely, by appeal to the results of lower-level severe tests that probe for specific sources of error. In the context of computer simulation modeling, this strategy would involve showing that standard sources of error that can arise in simulation studies either were absent from our simulation study or, if present and unable to be "subtracted out", were unlikely to have impacted the simulation results by more than a specified amount.

To implement this strategy, we would need an understanding of the canonical sources of error that can impact computer simulation results as well as procedures that severely test for the presence of those sources of error and/or reliably estimate the magnitude of their impact. Do we have either of these?

3.1 Sources of error in computer simulation studies

Scientists have formulated taxonomies of error sources that affect computer simulation results (e.g., Oberkampf et al. 1995; Roache 1998), and some philosophers have, too (e.g., Winsberg 1999b). Box 1 presents my own preliminary attempt at such a taxonomy.[1]

Study design error is typically an issue when the aim is to simulate repetitions of a procedure (as in the Monte Carlo simulation performed for the neutral currents experiments discussed by Mayo) or when multiple runs of a simulation model are being performed to explore the implications of uncertainty associated with the modeling assumptions. Both kinds of study involve sampling, and error can arise due to the limited number of simulation runs/trials performed or because the methods used to generate the sample of runs/trials are in some way inadequate. *Substantive modeling error* occurs when the equations or initial/boundary conditions chosen to represent

[1] My taxonomy draws upon but also differs somewhat from those offered by the authors mentioned above; there is not space to compare and contrast our taxonomies here.

Study Design Error
- Error due to limited number of simulation runs / trials
- Inadequate sampling method

Substantive Modeling Error
- Error in equations for modeled processes (form, parameter values)
- No representation of relevant processes
- Overly simplified/erroneous initial and/or boundary conditions

Data Processing Error
- Error introduced by processing of raw simulation results

Solution Algorithm Error
- Inapplicable solution algorithm
- Unstable solution algorithm

Numerical Error
- Discretization error
- Iterative convergence error
- Truncation error

Programming Error
- Inadequate/faulty program design
- Coding typo/mistake

Hardware-related Error
- Round-off error
- Internal malfunction
- External interference

Box 1 Sources of error in computer simulation studies

the target system are inappropriate, given the goals of the modeling study. Substantive modeling error includes errors of omission—cases in which relevant features of the target system are simply not given any representation in the model. *Data processing errors* are distortions or other errors that are introduced by the procedures (if any) used to process raw simulation data before they are delivered as simulation output/results. For instance, such processing might involve interpolating raw data from the regular grid of the simulation to points of interest that do not lie at the nodes of the grid (as when forecasts of weather conditions are generated for cities that lie between the points for which the model makes calculations).

Solution algorithm error occurs when a simulation study makes use of solution methods that are not capable of solving the equations of the model to the desired degree of accuracy. Closely related is *numerical error*, which occurs anytime the chosen solution algorithms employ numerical solution methods, since these methods deliver only approximate solutions to equations of interest; for these errors, the goal is usually to estimate and/or minimize their magnitude. *Programming error* occurs when the computer program in which the solution algorithms and processing procedures are embedded has not been properly designed or implemented. As model developers know

all too well, even the most minor of typos made when programming can significantly impact simulation results. Lastly, *hardware-related error* occurs if the computer that runs the program is not functioning properly or suffers some external interference, such as a power surge; in addition, some hardware-related error is always present in the form of round-off error, as a result of the finite storage precision of the computer.

3.2 Probing severely for error in computer simulation studies

While there is clearly plenty of opportunity for further analysis, the preceding discussion suggests that we do have some understanding of the different sources of error that can impact computer simulation results.[2] But are there procedures that we can use to severely test for their presence and/or to reliably estimate the magnitude of their impacts? While I cannot hope to provide a comprehensive answer to this question here, there are several remarks I would like to make.

First, as I will emphasize again in Sect. 4, explicit concern with severe testing for error is surprisingly rare in the context of computer simulation modeling. Nevertheless, something like severity considerations do sometimes appear. Examples can be found in discussion of simulations that involve estimating solutions to differential equations, especially in the area of computational fluid dynamics. For instance, Salari and Knupp (2000) and Roy (2005) describe a variety of tests for detecting coding mistakes that impact the order of accuracy of solutions and recommend the most "rigorous" and "sensitive" of the available tests. Although exactly what it means for a test to be rigorous or sensitive is not made explicit, Salari and Knupp say that when the recommended tests are passed "the probability of a coding mistake is deemed small" (2000, p. 18). Similarly, Roache (2002, pp. 9–10) claims that, for at least some types of equations, it is "highly unlikely" that a computer code passing one of these recommended tests would be "wrong", by which he seems to mean that it is highly unlikely that the code would contain errors that prevent it from solving the equations of the model to the order of accuracy that the chosen solution method, if implemented properly, is capable of delivering.[3]

Another example can be found in discussions of the reliability of proposed procedures for estimating the magnitude of errors that are known to be present in simulations that involve estimating solutions to differential equations. Roache (1994, 1997) describes a procedure for estimating the magnitude of discretization error—the error that occurs because solutions to continuous equations are being estimated on a spatial grid with finite rather than infinitesimal spaces between grid points. He indicates that his goal in developing this estimation procedure was to ensure that, in approximately 95% of the cases in which the procedure is used, the true discretization error will lie within the interval generated by the procedure (see Roach 2003, p. 731). From

[2] Salari and Knupp (2000), for instance, present a taxonomy of coding mistakes that could in principle be embedded within my taxonomy. Similar expansions might be performed for the other kinds of error.

[3] These recommended tests involve exercising the code on complex solution tasks and checking whether refinement of the spatial grid on which solutions are calculated leads to expected changes in the order of accuracy of the solutions; the expected changes are found via formal analysis of the exact solutions (see e.g., Roache 2002; Roy 2005).

the point of view of using simulation results as evidence, having this sort of reliable procedure for estimating the magnitude of discretization error would clearly be useful, since in order to sustain an argument concerning the likelihood that some simulation results would agree so well with a given hypothesis, were the hypothesis false, we need to take into account how sources of error known to be present are likely to have impacted the simulations results. Whether Roache's estimation procedure really does provide reliable information concerning the magnitude of discretization error, however, remains a topic of debate; he reports that in a suite of tests his procedure did achieve something like the 95% reliability for which he was aiming (Roache 2003), but other modelers have suggested that the results of those tests were insufficient for establishing confidence intervals, both because no underlying error distribution was specified and because the analysis nevertheless sometimes seemed to assume a normal distribution of errors, when there are reasons to think that such an assumption is unwarranted (see Wilson et al. 2004).

Second, I note that it remains to be seen whether formal statistical analysis can have as large a role (or the same role) to play in arguments for the evidential value of computer simulation results as Mayo suggests it has in arguments for the evidential value of results from traditional experiments.[4] Recall that, according to Mayo, one major use of formal statistical analysis in traditional experimental contexts is in developing arguments about the presence or absence of errors. Whether formal statistical models can be used this way in the context of computer simulation would seem to depend in part upon the nature of the sources of error that can impact simulation results—e.g., do coding mistakes of a certain type, discretization errors, etc. impact simulation results in patterns that can be captured with standard statistical models, such as Gaussian distributions? In the examples presented above, although the modelers were concerned with the reliability of their test procedures and collected data on the procedures' successes and failures in detecting particular errors, they did not investigate whether the observed distributions of errors could be assimilated to any particular formal models. (As just indicated, however, this issue did arise in critical discussion.) But even if it turns out that formal statistical analysis cannot play the same role as it does in the context of traditional experimentation, informal arguments from error might still be possible. For example, an experienced modeler might be prepared to argue informally that a problem known as numerical instability is unlikely to have significantly impacted simulation results of interest, because not only have all runs of the simulation model managed to complete (while numerical instability can cause a simulation to crash), but graphical displays of the simulation results have failed to show any of the other telltale signs that typically appear when numerical instability is present.

Third, I want to flag what seems to be a particularly difficult error-probing task, namely, that which involves testing for substantive modeling error. When thinking about this task, we should remember that what we are really interested in detecting is *problematic* substantive modeling error—i.e., substantive modeling error that renders the model inadequate for our purposes. We need not care if our simulation

[4] Berk et al. (2002) illustrate some of the roles that statistical concepts and tools might play in the evaluation of computer simulation models, but much work remains to be done in this regard.

model assumes that boundary conditions are rather simpler than they are in reality, for instance, as long as the model can still be used to provide evidence regarding hypotheses of interest about the target system. How difficult it is to devise severe tests for problematic substantive modeling error may be a function of, among other things, the extent to which we can intervene on and observe the real-world target system of interest. For example, if our simulation model reflects the assumption that an object has a particular shape or that a field has a particular structure, we may be able to test the accuracy of that assumption via traditional experimentation and/or observation and perhaps even to estimate via mathematical analysis how the results of our simulation are likely to be impacted by our assumptions being inaccurate in particular ways. I take the misspecification testing discussed by Mayo and Spanos (2004) to be closely related to, if not an instance of, this sort of testing of substantive model assumptions.

Lastly, I want to explore the possibility that we sometimes can circumvent the task of devising severe tests that specifically target substantive modeling error. The alternative approach envisioned here would involve establishing empirically the broader claim that particular kinds of results from simulation studies of a specified type (e-type results) can be used as reliable indicators of the truth/falsity of hypotheses of a particular sort (H-type hypotheses). Implementing this approach would require collecting data on how frequently H-type hypotheses are false when various degrees of fit between a pair consisting of an H-type hypothesis and an e-type result occur. Suppose we find that, when the degree of fit between e-type results and H-type hypotheses is at least as great as Ψ, the H-type hypotheses are only very rarely false. We might invoke this finding, along with evidence that a simulation study had been run in the usual way and without any new signs of error, to argue that e, the particular e-type simulation result generated in that study, constitutes good evidence for H, some particular H-type hypothesis whose degree of fit with e is at least as great as Ψ. Such an argument would be analogous to one that invoked empirical evidence of the high reliability of a type of home pregnancy test, along with evidence that the test was correctly deployed in the case at hand, in defending the conclusion that the test's delivering a result of "pregnant" constituted good evidence of pregnancy in the individual tested.[5]

Obviously, this approach could be implemented only if a number of requirements were met. A good estimate of the error statistics associated with the use of e-type results to test H-type hypotheses is needed, and obtaining this requires both that there exist a relatively stable distribution of errors from which we can sample and that we take an adequate sample from that distribution when we collect data on how frequently a given fit between an H-type hypothesis and an e-type result correctly indicates the truth/falsity of the H-type hypothesis.[6] In addition, even if we do have a good estimate

[5] The recognition that results of "pregnant" from a dozen different but reliable home pregnancy tests provides even stronger evidence of pregnancy raises interesting questions about the evidential value of agreement among results generated from several different computer simulation models. Though there is not space to address such questions in detail here, I note that the evidential value of such agreement in the case of simulation results will depend on the extent to which each of the models is reliable individually and on the extent to which the models' errors for such results are uncorrelated (just as in the case of the pregnancy tests).

[6] What counts as an adequate sample depends on several things, including the estimation procedure used, which may require a random sample from the error distribution.

of the relevant error statistics, we will be able to invoke it in arguing that some particular *e*-type results constitute good evidence for some particular *H*-type hypothesis only if the error statistics indicate that it is highly unlikely that there would be such a good fit between an *e*-type result and an erroneous *H*-type hypothesis.

Given these and other requirements, it is unclear how often this approach could be implemented in practice. There seems little prospect of doing so for model predictions of long-term climate change, for instance. Since climate predictions concern conditions decades from now, it would take a long, long time to collect statistics on how well the models perform on such prediction tasks (assuming some stable distribution of error exists—see below). Looking at model performance in simulating past climatic conditions will not suffice either, at least not for today's models; given that some observations of past climatic conditions have been used in constructing and tuning today's climate models, and recognizing the possibility that earth's climate is evolving, we cannot assume that errors in simulations of past climate are sampled from an error distribution identical to that associated with predictions of future climate. Indeed, if earth's climate is evolving, while our models are to some degree tuned to the climate of the recent past, then there may be no stable error distribution from which to sample!

On the other hand, it might be possible to implement this approach for short-term weather forecasts, since for these we can collect a significant amount of data on how well the models perform on particular prediction tasks of interest. We might check, for instance, how frequently over the last five years, when this simulation model predicted that the next day's high temperature at Chicago's O'Hare International Airport would be above average, the high temperature observed there on the next day turned out to actually be above average. Performance information like this for weather forecasting models is in fact collected and studied (e.g., Nachamkin (2004)) and is the sort of information that in principle might provide a basis for arguing for the evidential value of particular simulation results (e.g., that this model's predicting that the temperature tomorrow at O'Hare will be above average is good evidence that the temperature will be above average).

Still, there are many complications to be dealt with before this approach could be used even in the case of weather forecasting. For one thing, at most forecasting centers, the forecast procedure itself is constantly evolving; adjustments are made to the weather forecasting models and to the observing networks that supply data that are needed to initialize the models, suggesting that it would be a mistake to simply assume that data collected over a few years were sampled from a single error distribution. There is also the question of which class of predictions should be considered when using error statistics to argue for the evidential value of particular simulation results. For example, do we need error statistics for the class of predictions consisting of daily high temperatures in a particular temperature range for a single location in a single season of the year, or will statistics for the class consisting of predictions of daily high temperatures in any range for any of many locations in all seasons of the year suffice?[7] Presumably the answer depends in part on whether the error distributions associated with different prediction classes differ significantly from one another, but

[7] This issue is touched upon in Taylor and Leslie (2005).

for very narrowly-defined classes we may lack sufficient data to draw key conclusions about the distributions from which they are drawn. The importance of this question about the relevant prediction class should not be overlooked; in any given situation, the answer determines which hypotheses (if any) we can claim to test severely with particular simulation results.

4 Value of an error-statistical perspective on computer simulation modeling

Though much work remains to be done to understand the prospects for using simulation studies to severely test hypotheses of scientific interest, I contend that adopting an error-statistical perspective on computer simulation modeling would be valuable for at least two reasons: it would offer an interesting new way of thinking about computer simulation studies, and it would promote important and healthy changes in the practice of simulation model evaluation.

Adopting an error-statistical perspective would offer an interesting new way of thinking about computer simulation studies, namely, as procedures that may be capable of severely testing hypotheses of interest. On such a view, what is of fundamental interest (at least when it comes to providing evidence) is not whether a model can produce "realistic-looking" simulations of various real-world target systems, but rather the space of hypotheses about those target systems that can be severely tested using that model; simulation models are conceived of less as imitating or mimicking devices and more as putative hypothesis-testing tools. Instead of always focusing on the question of whether simulation results are true of a real-world target system, we can shift attention to the question of which range of hypotheses about that target system can be rejected or accepted in light of the production of those results by that model. For instance, instead of asking whether it is true that tomorrow's high temperature will be 70° F, as our weather model predicts, we might ask the slightly different question of which hypotheses about tomorrow's high temperature can be accepted or rejected in light of the production of a prediction of 70° F by our model.

Adopting an error-statistical perspective on computer simulation models also would promote important and healthy changes in the practice of simulation model evaluation. At present, this practice often lacks rigor and structure. Which tests or checks are performed is sometimes determined largely by convenience—how much time and computing power are available, past experience with evaluation techniques, the nature of the available visualization tools, etc. This would not be so worrisome if simulation models were being used only for heuristic purposes, but that is not the situation. In many cases, even when the ultimate aim is to use simulation models as evidential resources, model evaluation consists of little more than side-by-side comparisons of simulation output and observational data, with little or no explicit argumentation concerning what, if anything, these comparisons indicate about the capacity of the model to provide evidence for specific scientific hypotheses of interest. Moreover, conclusions drawn in such discussions often come in the form of qualitative judgments of the extent to which some set of simulation results correspond with observational data (e.g., "the simulation results agree reasonably well with the observations"), perhaps accompanied by some statement concerning the evaluator's "confidence" in the model

in light of that correspondence. This has largely been the situation, for instance, when it comes to the evaluation of computer simulation models used to project how earth's climate will change over the next century.[8]

An error-statistical perspective on computer simulation would provide some of the rigor and structure that is often missing from the practice of simulation evaluation at present. As discussed in the last section, an error-statistical approach would call for identifying potential sources of error and performing a series of tests designed to probe severely for the presence of those sources of error. Importantly, in order to claim that simulation results provide good evidence for some hypothesis of interest, we would be required to show that the potential sources of error were *unlikely* to have been present or to have impacted the results by more than a specified amount, rather than just that the evidence collected so far is *consistent* with their absence or their having minimal impact. With this requirement in mind, it becomes obvious that displaying side-by-side comparisons of model output and observational data is not enough—though the appearance of a good fit between model output and observational data is consistent with the absence of standard sources of error in simulation studies, it by no means establishes that such sources of error are unlikely to be present; for one thing, issues such as model-data dependence (which is common in the context of climate modeling, for instance) must be addressed. An error-statistical perspective, with its explicit focus on the nature of the test procedure, pushes us to confront such issues.

Such a perspective could be expected to have additional, related benefits as well. It would work against overconfidence in simulation results, not only by setting a demanding standard for good evidence but also by forcing us to consider what we know (and don't know!) about the impacts of potential sources of error on our simulation results; in the process, we may come to realize either that there are important sources of error for which we have not yet probed at all or that some sources of error are indeed impacting our simulations, despite our having judged the results to look realistic. Such realizations are particularly valuable, not least because they can direct and focus our attempts to improve our simulation models. In addition, since an error-statistical perspective would encourage us to formulate not just any old tests, but tests that probe severely for error, we might be less likely to waste our limited resources on tests of our models that are actually rather uninformative. The same focus on severe testing might also lead us to take more seriously issues of simulation study design, which are often overlooked at present but which impact what we can even possibly claim to have learned from our simulations.

5 Concluding remarks

The foregoing is a preliminary discussion of how Mayo's error-statistical framework might be applied to address important questions about the evidential status of computer

[8] Admittedly, evaluation of climate models is very difficult, for a variety of reasons. Also, of late there is growing recognition of the need to consider what side-by-side comparisons of climate model output with observational data really indicate, if anything, about climate models as predictive tools (see IPCC 2007, Chap. 8). Still, there remains much room for improvement when it comes to the practice of climate model evaluation.

simulation results. It is clear that much work remains to be done to flesh out, and even to determine the real prospects for, an error-statistical epistemology of simulation. It is possible that further analysis will reveal that, given the stringent requirements of the error-statistical account of evidence, we rarely are warranted in taking simulation results to be good evidence for hypotheses of the sort that typically interest us in science. I have not attempted to make any arguments about the likelihood of such an outcome. Regardless, it seems better that we work hard to identify the capacities and limitations of our models as evidential resources, as the error-statistical approach with its emphasis on severe testing prompts us to do, than that we simply hope that a "reasonably good fit" between model output and observational data indicates that our models can be trusted to tell us what we want to know.

Acknowledgements Thanks to Deborah Mayo, Kent Staley, Phil Ehrlich, Lenny Smith and anonymous referees for very helpful suggestions on earlier drafts of the paper. Thanks also to participants at ERROR06 for encouraging feedback and stimulating discussion. Some of the research for this paper was conducted when the author held a postdoctoral fellowship supported by a National Science Foundation Research and Training Grant in "Proof, Persuasion, and Policy" (SES-0349956).

References

Beck, B. (2002). Model evaluation and performance. In A. H. El-Shaarawi & W. W. Piegorsch (Eds.), *Encyclopedia of environmetrics* (pp. 1275–1279). Wiley: Chichester.

Berk, R. A., Bickel, P., Campbell, K., Fovell, R., Keller-McNulty, S., Kelly, E., Linn, R., Park, B., Perelson, A., Rouphail, N., Sacks, J., & Schoenberg, F. (2002). Workshop on statistical approaches for the evaluation of complex computer models. *Statistical Science, 17*, 173–192.

IPCC (2007). In S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, & H. L. Miller (Eds.), *Climate change 2007: The physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge: Cambridge University Press.

Mayo, D. (1996). *Error and the growth of experimental knowledge.* Chicago: University of Chicago Press.

Mayo, D. (2000). Experimental practice and an error statistical account of evidence. *Philosophy of Science, 67*(Supp), S193–S207.

Mayo, D. (2005). Evidence as passing severe tests: Highly probable vs. highly probed hypotheses. In P. Achinstein (Ed.), *Scientific evidence: Philosophical theories and applications* (pp. 95–128). Baltimore: Johns Hopkins.

Mayo, D., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science, 71*, 1007–1025.

Nachamkin, J. E. (2004). Mesoscale verification using meteorological composites. *Monthly Weather Review, 132*, 941–955.

Oberkampf, W. L., Blottner, F. G., & Aeschliman, D. P. (1995). Methodology for computational fluid dynamics code verification/validation. AIAA Paper 1995–2226.

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

Roache, P. J. (1994). Perspective: A method for uniform reporting of grid refinement studies. *Journal of Fluids Engineering, 116*, 405–413.

Roache, P. J. (1997). Quantification of uncertainty in computational fluid dynamics. *Annual Review of Fluid Mechanics, 29*, 123–160.

Roache, P. J. (1998). *Verification and validation in computational science and engineering*. Albuquerque: Hermosa.

Roache, P. J. (2002). Code verification by the method of manufactured solutions. *Journal of Fluids Engineering, 124*(1), 4–10.

Roache, P. J. (2003). Conservatism of the grid convergence index in finite volume computations on steady-state fluid flow and heat transfer. *Journal of Fluids Engineering, 125*, 731–732.

Roy, C. J. (2005). Review of code and solution verification procedures for computational simulation. *Journal of Computational Physics, 205*, 131–156.

Rykiel, E. J. (1996). Testing ecological models: The meaning of validation. *Ecological Modelling, 90*, 229–244.

Salari, K., & Knupp, P. (2000). Code verification by the method of manufactured solutions. Sandia Report SAND2000-1444.

Taylor, A. A., & Leslie, L. M. (2005). A single-station approach to model output statistics temperature forecast error assessment. *Weather and Forecasting, 20*(6), 1006–1020.

Wilson, R., Shao, J., & Stern, F. (2004). Discussion: Criticisms of the "Correction Factor" verification method. *Journal of Fluids Engineering, 126*, 704–706.

Winsberg, E. (1999a). Sanctioning models: The epistemology of simulation. *Science in Context, 12*, 275–292.

Winsberg, E. (1999b). Simulation and the philosophy of science: Computationally intensive studies of complex physical systems. Dissertation, Indiana University.