

Computer Simulation Modeling Through an Error-Statistical Lens

Wendy S. Parker

Abstract. When do computer simulation studies provide good evidence for hypotheses about real-world target systems? On what basis can scientists argue that particular simulation results constitute good evidence for such hypothesis? In this paper, I consider how Deborah Mayo’s error-statistical epistemology of science might be employed to help answer questions like these about the evidential status of computer simulation results. I also discuss some advantages that an error-statistical approach to the epistemology of computer simulation would have over the confidence-building framework that I take to be common among scientists at present.

1. Introduction

Over the past several decades, computer simulation modeling has emerged as an important mode of research in many, if not most, scientific fields. Painting with a broad brush, we can identify at least two kinds of 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 new hypotheses and explanations worthy of further investigation via observation and experiment. There seems to be broad agreement among scientists and philosophers alike that computer simulation models often have this heuristic value. Second, computer simulation models might serve as evidential resources, i.e. they might be used in investigations that, like traditional experiments, are meant to provide evidence for hypotheses about the natural world. Surveying actual modeling studies, it seems clear that scientists sometimes do undertake computer simulation studies with the aim of producing evidence for hypotheses about real-world target systems. The goal of many climate change simulations, for instance, is to provide accurate estimates of how much the average temperature of Earth’s atmosphere will increase in response to increased greenhouse gas emissions. Nevertheless, the idea that computer simulation results can count as good evidence for hypotheses about real-world systems is far from universally accepted. And even among scientists who don’t find the idea contentious, the question of when it is appropriate to take simulation results to be good evidence for real-world hypotheses remains a topic of considerable debate and concern.

In this paper, I consider how Deborah Mayo's error-statistical epistemology of science (Mayo 1996, 2000) might be employed to clarify and help answer questions about the evidential status of computer simulation results. I draw upon her account to help illuminate the following questions in particular: When do computer simulation studies provide good evidence for hypotheses about real-world target systems? On what basis can scientists argue that particular simulation results constitute good evidence for some hypothesis? In Section 2, I introduce Mayo's error-statistical epistemology of science, which focuses on experimental inquiry, and explain how computer simulation has figured in her account. In Section 3, I apply the core ideas of Mayo's epistemology to the practice of computer simulation modeling. Doing so delivers answers to my questions about the evidential status of computer simulation results and provides helpful structure and direction for future work on the topic. In Section 4, I discuss some advantages that an error-statistical approach to the epistemology of computer simulation would have over an approach that I take to be common among scientists at present. Finally, in Section 5, I offer some concluding remarks.

2. Mayo's Error-Statistical Epistemology of Science

In recent works, Deborah Mayo has developed an account of how experimental inquiry can provide good evidence for scientific hypotheses (see Mayo 1996, 2000). In a nutshell, it is because experiments, if they are designed and carried out right, can provide severe tests of hypotheses. A *severe test* of some hypothesis H is a procedure that has a high probability of rejecting H , just in case H is false. The higher the probability that a test will reject H , if and only if H is false, the more severe the test of H . We say that H *passes* a severe test (with some data e), just in case it is very unlikely that e would have accorded so well with H , if H is in fact false. If H passes a severe test with e , then e constitute *good evidence for*—or *a good indication of*— H . The more severe the test passed by H with e , the better evidence e is for H (see Mayo 1996, Ch.6, and 2000, p.S198 for more details related to these definitions and requirements).

On Mayo's analysis of experimental inquiry, severe testing occurs on multiple levels. On the largest scale, we aim for the experiment itself to provide a severe test of some primary hypothesis that interests us. Suppose, for example, we hypothesize that a particular drug dramatically reduces the rate at which certain cancers spread in the body, and we conduct an experiment to test this hypothesis. What we want, says Mayo, is for our experiment to have a

high probability of indicating that the drug makes no difference, just in case it in fact makes no difference. Such an experiment would be a severe test of our hypothesis about the drug's efficacy. But whether any experiment that we actually conduct really is a severe test of some hypothesis of interest depends on how the experiment is designed and executed. Thus, a second, lower level of severe testing must probe for error in the design and execution of the experiment. Error might arise for any number of reasons: the design assumptions of the experiment were not met, the experimental apparatus malfunctioned, the data processing techniques were biased, a confounding factor was inadequately controlled, etc. In light of these potential sources of experimental error, we must conduct a battery of severe tests, each designed to probe for the presence of one or more errors. Ideally, we want to show that it is very unlikely that any of the plausible sources of error were present. But even if we determine that some of them are present—indeed, we may have every reason to expect that some errors, such as random measurement error, are present—we may be able to argue either that it is very unlikely that they impacted the results by more than a specified amount or that their impact on the results can be subtracted out. Only after we carry out this lower-level probing for error in the experiment itself do we have warrant for accepting the results of the experiment and thus for accepting or rejecting in light of those results our initial, primary hypothesis about the drug's efficacy. In other words, even if the results of the experiment seem to indicate a huge difference between the cancer progression in the treatment and control groups, on Mayo's view we cannot take this to be good evidence for our hypothesis about the drug's efficacy until we can argue that the observed difference is not a consequence of some error in the design or execution of our experiment.

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 experimental error, not least because they can help us to model error—to determine what we are more and less likely to observe when we carry out a given test procedure, if a particular source of error is present (1996, p.164). With this information, we can determine how likely it is that we would have observed what we actually observed, if the error were present. If it is very unlikely that we would have observed what we actually observed, then we can argue that the source of error was unlikely to have been present. Of course statistical tools don't give us

answers “from thin air” (ibid, 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 ibid, pp.549-462). Perhaps the most familiar use of statistical tools and techniques in this way is in the design and analysis of randomized clinical trials, where they are used (among other ways) to decide on the sizes of the treatment and control groups and to attach significance levels to results concerning observed differences between those groups, but Mayo also discusses examples from other scientific domains, including physics (e.g. ibid, pp.92-99). 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. On the contrary, sometimes more qualitative arguments about what it would (and wouldn’t) be like if a particular source of error were present in a simulation can be perfectly appropriate—what Mayo refers to as “informal” arguments from error.

It is in discussing the need to model experimental error that Mayo makes reference to computer simulation studies. As she illustrates, scientists sometimes carry out computer simulations to help them estimate what it would be like if some source of error were present in an experiment (see e.g. her neutral currents example, in 1996, p.163). Interestingly, Mayo’s acceptance of the use of simulations to model error in this way seems to commit her to the view that computer simulations can sometimes provide good evidence for hypotheses about the real-world systems they represent. For presumably, simulation results could only be used in the way Mayo suggests—i.e. to argue that some source of error was absent from an experiment—if those results were a good indication of the likely effects of that source of error in that experiment. That is, simulation results could only be used in the way Mayo suggests if they constituted good evidence for an hypothesis about what it would be like if that source of error were present in that experiment. Such an hypothesis is an hypothesis about a real-world system, namely, the particular system that constitutes the experimental set-up. Strangely enough, then, by considering Mayo’s account of the epistemology of experimental inquiry, we are led back to our original questions about the evidential status of computer simulation results. It appears that as we draw upon core ideas from Mayo’s account to address those questions in the next section, we will at the same time be filling in some further details of her error-statistical epistemology of experiment!

3. Computer Simulation through an Error-Statistical Lens

3.1 Requirements for good evidence

How can Mayo's error-statistical account help us to address our questions about the evidential status of computer simulation results? For starters, and most obviously, she gives us an account of what it means for data to be good evidence for an hypothesis. We can apply this directly in the context of computer simulation modeling, taking selected results of the simulation to be the data. Consider our first question: When do computer simulation studies provide good evidence for hypotheses about real-world target systems? A necessary condition is that the simulation study be a severe test of the hypothesis, that is, that the simulation study have a high probability of revealing H to be in error, if and only if H is in fact in error. In addition, H must pass the severe test afforded by the simulation study. Putting the answer to our question more sharply:

- Computer simulation studies provide good evidence (in the form of simulation results) for some hypothesis H about a real-world target system to the extent that it is unlikely that the simulation results would have accorded with H as well as they actually did, if H is false.

Reflecting on this, we are immediately led to our second question: On what basis can scientists argue that some particular simulation results actually do constitute good evidence for some hypothesis? That is, on what basis can scientists claim that it is unlikely that the simulation results would have accorded with H as well as they actually did, if H is false? If we think of simulation studies as test procedures and then follow Mayo's lead, we will answer this with an appeal to lower-level probing for error. As with the canonical sources of error that arise in the context of experimental inquiry, scientists will need to probe for standard sources of error that might have impacted the simulation results, before they can claim that the simulation study provided a severe test of the hypothesis of interest. They will need to show that these sources of error either are absent from the simulation or, for errors known to be present, are unlikely to have impacted the results by more than a specified amount. Only then will scientists have warrant for accepting or rejecting in light of the simulation results the primary hypothesis about the system being modeled.

Of course, for scientists to carry out this lower-level testing, they need to have some idea of the standard types of error that arise in the context of computer simulation. What are the ways in which computer simulation studies can go wrong? Scientists concerned with the evaluation of computer simulation models have already attempted to formulate taxonomies of error (see Oberkampff et al. 1995, Roache 1998, NPARC 2005). Some philosophers have, too (see Winsberg 1999). Table 1 presents my own error taxonomy for computer simulations studies.^{1, 2}

Hardware-related Error
--Round-off error
--Internal malfunction
--External interference
Algorithm Error
--Faulty design of solution algorithm
Programming Error
--Faulty program design
--Coding mistake/typo
Numerical Error
--Discretization error (time and space)
--Iterative convergence error
--Truncation error
Substantive Modeling Error
--Error in substantive modeling assumptions (equations, constants)
--Omission of relevant processes
--Overly simplified/erroneous initial conditions
--Overly simplified/erroneous boundary conditions

Table 1. Sources of Error in Computer Simulation Modeling

¹ My taxonomy differs somewhat from the those offered by the authors mentioned above, but I will not take time to discuss the differences here.

² There may be important sources of error that I have overlooked when constructing my taxonomy. I welcome advice on improving and/or expanding Table 1.

The sources of error in Table 1 ultimately have to do with:

- (1) whether the computer that performs the calculations is functioning properly [internal malfunction, external interference],
- (2) whether the algorithm chosen is capable of solving the equations of the simulation model to the intended degree of accuracy [algorithm error] and is implemented properly [programming errors]
- (3) whether the equations and initial/boundary conditions chosen to represent the system are appropriate, given the goals of the modeling task [substantive modeling errors],
- (4) whether the simulation uses numerical solution methods [numerical errors], and
- (5) the finite storage precisions of the digital computer [round-off error].

As in the context of experimental inquiry, we sometimes may be able to design single tests that probe severely for several of these potential sources of error at once, while for other potential sources of error it may be that only a number of tests together can be considered a severe test for that error.

Let us survey the progress that has been made. We now have relatively clear and precise statement of what is required for simulation results to provide good evidence for some real-world hypothesis. We also have a sketch of a taxonomy of error, which helps to make more concrete what the requirements for good evidence amount to in the context of computer simulation. With this progress, we have put ourselves in a better position to formulate and address other questions about the evidential status of simulation results, whether in particular cases or in general.

3.2 Meeting the requirements for good evidence

An obvious question to ask, for example, is whether computer simulation studies ever actually meet the requirements for providing good evidence for hypotheses of interest. Given the foregoing discussion, we can reformulate this as a question related to the errors in Table 1: Are there procedures that we can use to severely test for and/or estimate the magnitude of these errors? Although this question is of utmost importance from an error-statistical point of view, I will not attempt to give a complete answer here. However, I will make a few remarks on the matter before moving on to a discussion of what I perceive to be some important benefits of looking at computer simulation modeling through an error-statistical lens.

First, I want to give some sense of what severity considerations look like in current simulation practice, in some of the rare cases in which modelers explicitly discuss strategies for probing for error in simulations. When it comes to programming errors, Roy (2005) emphasizes in a recent review article that some test procedures are more “sensitive” to programming errors than others and recommends using the most sensitive of the ones he considers. This procedure involves exercising the code with complex solution tasks and checking to see that refinement of the spatial grid on which the solutions are calculated leads to expected changes in the accuracy of the solutions (for technical details, see Roy 2005 and Roache 2002). This test procedure has itself been tested by Knupp and Salari (2003), who wanted to see how reliably it could detect programming errors. They conducted a blind study in which an investigator used the test procedure to try to determine whether a simulation model contained a programming mistake. They found that the test procedure succeeded in detecting the presence of each of a dozen different programming errors and that it correctly reported no error on the single occasion in their study when none had been introduced (see *ibid*). Though there were additional coding errors that the test procedure failed to detect, this was because these errors turned out not to impact the accuracy of the solutions found and so were of less concern.³

Similarly, for errors known to be present in numerical simulations, procedures recently have been developed for estimating their likely magnitude. For instance, Roache (1994, 1997) has developed a procedure for estimating the magnitude of discretization error—the error due to the model equations’ being solved on a (virtual) spatial grid with finite rather than infinitesimal spaces between grid points. Again, the procedure involves checking the rate at which solutions converge in response to refinement of the grid, but an additional “factor of safety” multiplier is relied upon in generating the estimation of the discretization error (see *ibid*). The factor of safety has been chosen with the explicit goal of ensuring that, in about 95% of the cases in which the estimation procedure is used, the true discretization error lies within the error interval generated by the procedure (see *ibid* for details). Interestingly, Roache seems to have selected his factor of safety values on the basis of educated guesswork,

³ By contrast, another test procedure mentioned by Roy (2005) involves simply asking experts whether the solutions generated by the simulation code appear accurate. Roy refers to this as a “less rigorous” test, but it seems likely that it would also be of low severity, at least for simulation models that incorporate numerous and complicated equations.

rather than theoretical analysis of the mathematics. In support of his selections, he reports that in ongoing testing his procedure is achieving something like the 95% reliability for which he was aiming (Roache 2003). Modelers with rival estimation procedures, however, invoke statistical considerations to deny that Roache has established the reliability of his procedure (see Wilson et al. 2004), and debate is ongoing.

From the fact that a few modelers seem concerned with something like the severity of some error-probing procedures, however, we should not conclude that the evaluation of simulation models is in general focused on careful and severe probing for error. As I will discuss in the next section, simulation evaluation is more commonly framed as a confidence-building activity, rather than one concerned with severe testing.

Second, I want to point out that it remains to be seen whether formal statistical analysis can have as large a role (or the same role) to play in the epistemology of computer simulation as Mayo suggests it has in the epistemology of experiment. Recall that, according to Mayo, one major use of formal statistical analysis in experimental contexts is in the modeling of error, i.e. in the determining of which experimental outcomes are more and less likely if an error is present. Whether formal statistical models can be used this way in the context of computer simulation would seem to depend upon the nature of the errors—do coding mistakes of a certain type impact simulation results in a pattern that can be captured with a standard statistical model (e.g. a Gaussian distribution)? In the examples presented above, although the modelers were concerned with the reliability of their test procedures, they did not investigate whether the observed distribution of error could be assimilated to any particular formal model.

Third, I want to flag what seems to be a particularly difficult error-probing task, namely that which deals with substantive modeling error. Is there any hope of devising severe tests for substantive modeling errors? When thinking about this question, we should keep in mind that what we are really interested in detecting is substantive modeling error that will undermine our achieving the goals of our simulation study. We need not care if our simulation relies on simplified boundary conditions, for example, as long as the simplification does not prevent us from simulating accurately enough those features of the target system that matter, given what we want to learn about that system. So the real question is whether we can devise severe tests for substantive modeling errors that matter. How difficult it is to devise such tests

may be a function of, among other things, the particular goals of the simulation study and our ability to easily intervene on and observe the real-world target system; it may be relatively easy if the goal of the simulation is to make accurate predictions of a readily observable system (e.g. predictions of the occurrence of rain in Chicago during the following day), but virtually impossible if the goal is to explain of the behavior of a complex and inaccessible system.

When it comes to what seem to be the relatively easy cases, such as those concerning the prediction of rain in Chicago on the following day, one way of overcoming the challenge of ruling out substantive model error comes immediately to mind. It involves checking whether the prediction turns out to be accurate, when compared with observations.⁴ If it does, and if we can rule out or otherwise bound all the other plausible sources of error in our simulation, then the fact that the prediction was accurate constitutes a demonstration of the absence of substantive modeling error (that matters) in that particular simulation. For purposes of prediction, of course, this is a rather special and relatively uninteresting case. Without additional argument, we cannot take the accurate prediction to be a good indication of the absence of substantive modeling error (that matters) when it comes to other simulation studies we might conduct with our model, including studies undertaken to predict rain in Chicago on other days. And we might say that our rain prediction model isn't of much value if we can only say that one of its predictions constitutes good evidence of rain after we've checked whether it actually rained. In any case, looking beyond this example, further investigation of the prospects for severe testing of substantive model error is clearly needed.

4. Advantages of Viewing Computer Simulation through an Error-Statistical Lens

Although much work remains to be done in filling in the details of an error-statistical epistemology of computer simulation, I contend that adopting an error-statistical approach would promote some important and healthy changes in how scientists and philosophers think about computer simulation modeling and in the practice of simulation model evaluation.

At present, the practice of simulation model evaluation often lacks rigor and structure. Which tests or checks are performed is often determined largely by convenience—how much time and computing power are available, past experience with evaluation techniques, the

⁴ I am assuming here that we have probed severely for error in the process that produced the observations we use to judge the accuracy of the simulation's prediction.

nature of the available visualization tools, etc. In many cases, little or no attention seems to be paid to what, if anything, these tests are really capable of telling us about the capacity of the model to provide good evidence for hypotheses of interest. If a simulation model produces results that accord well with available data on past values taken by parameter X, does this tell us anything about the capacity of the model to predict future values of parameter Y? At present, questions like these are rarely explicitly addressed in discussions of model quality. Too often, all that is reported in such discussions is the extent to which the simulation results fit with some set of observational data, with that fit pointed to as somehow indicative of the general quality of the model. Even among modelers who clearly do recognize the importance of probing for error and estimating the uncertainty associated with simulation results, model evaluation is commonly framed as a confidence-building activity (e.g. Roache 1998; Roy 2005), through which we gradually accumulate evidence for the adequacy of a simulation model for a range of tasks (e.g. Oberkampf and Trucano 2002, p.26).

An error-statistical approach to the epistemology of simulation would provide some of the rigor and structure that is currently missing from the practice of simulation model evaluation. As we saw above, an error-statistical approach would require identifying sources of error and performing a series of tests designed to show that those errors are unlikely to be 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.

In addition, from an error-statistical perspective, we would be led to think of simulation modeling studies themselves in a new way, namely, as procedures that may be capable of providing severe tests of hypotheses of interest. If we think of modeling studies in this way, we no longer fixate on how closely simulation output fits (or can be made to fit) with available observational data. Rather, when confronted with a simulation model, what is of fundamental interest (at least when it comes to providing evidence) is the space of hypotheses that can be severely testing using that model. We don't focus on how much confidence we should have in a particular simulation result, but rather on which range of hypotheses can be rejected or accepted in light of the production of that result by that model.

Such a shift in our perspective on models can be expected to have important practical benefits. For one, it would work against overconfidence in simulation results. This is both because of the high standard that the error-statistical view sets for simulation results to

constitute good evidence and because it forces 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 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, because they can direct and focus our attempts to improve our simulation models and to conduct informative (i.e. highly severe) tests of those models. More generally, because we do aim to formulate tests that probe severely for error, we may be less likely to waste our limited model-testing resources on tests that are actually not very informative.

5. Concluding Remarks

[I will comment on the possibility that we rarely can have warrant for taking simulation results to be good evidence for real-world hypotheses that we actually care about. I have not attempted to make any arguments about the likelihood of such a state of affairs. I will suggest that, in any case, it is better that we recognize (and work to identify) the evidential limitations of our models.]

References

Knupp, P. and K. Salari (2003) *Verification of Computer Codes in Computational Science and Engineering*. Boca Raton: Chapman and Hall/CRC.

Mayo, D. (1996) *Error and the Growth of Experimental Knowledge*. Chicago: Univ. of Chicago Press.

--- (2000) “Experimental Practice and an Error Statistical Account of Evidence”, *Philosophy of Science* 67(Supp):S193-S207.

NPARC (2005) “Uncertainty and Error in CFD Simulations”, NPARC Alliance Verification and Validation Web Site, <<http://www.grc.nasa.gov/WWW/wind/valid/tutorial/errors.html>>, accessed November 7, 2005.

Oberkampf, W.L., Blottner, F.G. and D.P. Aeschliman (1995) “Methodology for Computational Fluid Dynamics Code Verification/Validation”, AIAA 1995-2226.

Oberkampf, W. L. and T.G. Trucano (2002) “Verification and Validation in Computational Fluid Dynamics”, Sandia National Laboratory Report, SAND2002-059.

Roache, P. J. (1994) “Perspective: A Method for Uniform Reporting of Grid Refinement Studies,” *Journal of Fluids Engineering* 116: 405–413.

--- (1997) “Quantification of Uncertainty in Computational Fluid Dynamics”, *Annual Review of Fluid Mechanics* 29:123-160.

--- (1998) *Verification and Validation in Computational Science and Engineering*. Albuquerque: Hermosa.

--- (2002) “Code Verification by the Method of Manufactured Solutions”, *Journal of Fluids Engineering* 124(1): 4-10.

--- (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.

Wilson, R. et al (2004) “Discussion: Criticisms of the ‘‘Correction Factor’’ Verification Method”, *Journal of Fluids Engineering* 126:704-706.

Winsberg, E. (1999) *Simulation and the Philosophy of Science: Computationally Intensive Studies of Complex Physical Systems*. Ph.D. Dissertation. Bloomington: Indiana University.