How to think about models and their evaluation – a proposal

Wendy Parker
Department of Philosophy, Ohio University
parkerw@ohio.edu

In this talk…

I. Scientific models as representational tools

- conceptualizing model quality and model evaluation

II. The practice of model evaluation

- from confirmation to severe testing

I. Scientific models as representational tools

Scientific models as representational tools...

- Scientific models, including complex computer models, are representations – entities that stand for, or stand in for, other entities.

- They are also tools – entities used by scientists to help them accomplish one or more aims, whether epistemic (prediction, explanation) or practical (communication, control).

Separating some questions about models…

- What makes some entity a model? In virtue of what is one entity a model of another?
  - …just someone's intention/decision that the first entity should stand for, or stand in for, the other.
- Why do scientists choose some entities rather than others to be models of a given target system?
  - …an empirical question; often in part because the chosen entities are thought to share relevant properties with the target, where relevance is a function of the aims of the modeling study.
- When is a scientific model a good one?

Model quality as purpose-relative…

- A model is a good model to the extent that it is adequate for the purpose(s) for which scientists want to use it.
  - Approximations, simplifications, false assumptions are not shortcomings of a model if, despite them, the model remains adequate for the intended/chosen purpose(s).
  - Something might be a good model (for purposes P) even though it incorporates a number of false or even meaningless assumptions.
- Judgments of model quality should be made relative to a set of purposes *P*, rather than simpliciter.
  - "M is a good model of X for P" rather than just "M is a good model of X"

Model evaluation as purpose-relative…

- Model evaluation is an activity aimed to determine the extent to which a model is adequate for a given set of purposes.
  - Might involve investigating the accuracy of modeling assumptions, but it need not always (e.g. could focus on predictive performance alone)

- Put differently, model evaluation is the activity of seeking evidence of the adequacy or inadequacy of a model for a given set of purposes.

Contrast with a view ...

- ... according to which model quality is a matter of the truth of modeling assumptions and the accuracy of all model output
  - so that approximations, simplifications, false assumptions, and inaccuracies in output are always shortcomings of a model, reducing its quality, regardless of the purposes for which the model is used.
- ...and model evaluation is concerned with
  - collecting evidence that the modeling assumptions are true, or
  - determining whether the modeling assumptions are true / false.

Testing and truth

- Oreskes et al (1994) reminded us that even if a model passes a battery of tests against observations, this is not a demonstration of the truth of the modeling assumptions.
  - ptolemaic astronomy, phlogiston theory
- This point may have been taken to heart by many, but model evaluation is still often conceptualized as fundamentally about investigating the truth/falsity of modeling assumptions, rather than the adequacy of the model for a given set of purposes.

Testing and truth (cont'd)

- "Science today recognizes that there is no way to prove the absolute truth of any hypothesis or model, since it is always possible that a different explanation might account for the same observations. …Rather, the test should be whether a theory or model is false." (IPCC 1997, 8)

- This seems a bizarre way to think about model evaluation, given that we almost always start out with very good reason to think that some of the modeling assumptions are not true of the system being modeled.

Testing and truth (cont'd)

- "It is always possible to find errors in simulations of particular variables or processes in a climate model. What is important to establish is whether such errors make a given model "unusable" in answering specific questions." (IPCC 2001, 474).

- Nevertheless, this is followed by:

  "It is important to remember that the types of error measurement that have been discussed are restricted to relatively few variables. It has proved elusive to derive a fully comprehensive multi-dimensional "figure of merit" for climate models."

II. The practice of model evaluation

Confirmation and truth

- Oreskes et al (1994) also…
  - warned that terms like "verification" and "validation" carried connotations of establishing the truth of modeling assumptions and
  - suggested that talk of model "confirmation" would be more appropriate in the context of model evaluation.
- A model or theory is confirmed when a prediction derived from it (+ initial and boundary conditions) differs from observations by less than observational uncertainty.
- Typically, instances of confirmation are viewed as somehow accumulating evidence for, but not establishing, the truth of the assumptions/hypotheses from which the predictions were derived.

Confirmation of what?

- But given a view of models as representational tools, and the prior recognition that our models do incorporate false assumptions, what is it that we'd want to confirm?
- Not the modeling assumptions, i.e. not the hypothesis that the modeling assumptions are true.
- But rather the hypothesis that the model is adequate for the chosen purpose(s).

Confirmation of what? (cont'd)

- We want to ask: If the model were adequate for this purpose, what sorts of predictive (or retrodictive) successes should it have? ...and then check whether the model achieves those successes.

- But "successes" on the purpose-relative account often will not require accuracy to within observational uncertainty nor complete absence of failed predictions over many trials, contrary to the situation when it is the truth of the assumptions that we mean to investigate.

An advantage…

- When facing a new modeling task, we must ask again: what are the expected successes if the model is adequate for this purpose?
- We can't simply assume that the other successes had by the model constitute good evidence of its adequacy for the new purpose:
  - perhaps the new task is more demanding (i.e. we expect various predictions to be much more accurate if the model is adequate for the new purpose)
  - or perhaps if the model is adequate for the new task, we expect that when predictions go wrong, they will go wrong in a different way.
- This unwarranted assumption is perhaps encouraged when model evaluation is thought of as activity that provides evidence for the truth of assumptions or general/overall model quality, but it is discouraged when evaluation is

Beyond confirmation…severe testing

- Perhaps another concept would help us to further improve the practice of model evaluation.
- On a confirmation-driven approach, we aim to accumulate evidence for the adequacy of a model for a set of purposes, but our attention is not necessarily focused on testing our models in the most informative ways.
- We may favor tests of convenience, because they too may be able to help us in our quest to accumulate the desired evidence, even if they are unlikely to reveal to us that our model is inadequate for the purpose of interest (if in fact it is inadequate).

Beyond confirmation…severe testing (cont'd)

- Perhaps it would be useful to approach model evaluation as an activity aimed to severely test, rather than confirm, the hypothesis that the model is adequate for the purposes at hand.

- A severe test of some hypothesis *H* is a procedure that is likely to indicate that *H* is false, if and only if *H* is in fact false (Mayo 1996, 2000).

Beyond confirmation…severe testing (cont'd)

- I will not make claims about whether, when, how exactly severe testing for adequacy-for-a-purpose might be achievable, but I expect that the procedure generally won't involve a single comparison of model output with observations.
- I can imagine approaches that involve either…
  - a series of comparisons of model output with relevant observations
  - and/or a collection of lower-level severe tests for particular errors that we have reason to think might undermine the adequacy of the model for the desired purpose.

A preliminary taxonomy of errors in simulation studies

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

--Purpose-relevant error in substantive modeling assumptions (equations, constants)
--Omission of purpose-relevant processes
--Overly simplified/erroneous initial conditions (given purpose)
--Overly simplified/erroneous boundary conditions (given purpose)

Advantages of a severe testing approach...

- I suggest that it would be advantageous to keep the idea of severe testing in mind, at least as a kind of regulative ideal, when approaching the task of model evaluation.

- When considering alternative test procedures we might perform, we will be led to consider which will be more likely to reveal model inadequacy (for our purposes) if such inadequacy is present.
  - away from tests of convenience if lacking in severity relative to others
  - perhaps toward better use of limited model evaluation resources

Advantages of a severe testing approach (cont'd)...

- Likewise, when presented with results of a test of a model (e.g. that there were such-and-such successful predictions), we won't just cheer, but will be led immediately to ask further questions about the test procedure and its ability to indicate inadequacy-for-purpose.
- For instance, we will consider issues of model/data dependence resulting in artificial inflation of model/data agreement:
  - Prior tuning of model against some of the test data
  - Development of parameterizations using some of the test data
  - "Data" that were processed using models with assumptions in common with the model under test (as in reanalysis data in climate change)