A Hierarchical Systems Approach to Model Validation

Steve Easterbrook
Dept of Computer Science, University of Toronto
sme@cs.toronto.edu

> All models are wrong, but some are useful
> - George Box

Introduction

Discussions of how climate models should be evaluated tend to rely on either philosophical arguments about the status of models as scientific tools, or on empirical arguments about how well runs from a given model match observational data. These lead to quantitative measures expressed in terms of model bias or forecast skill, and ensemble approaches where models are assessed according to the extent to which the ensemble brackets the observational data.

Such approaches focus the evaluation on *models* per se (or more specifically, on the simulation runs they produce), as if the models can be isolated from their context. Such approaches may overlook a number of important aspects of the use of climate models:

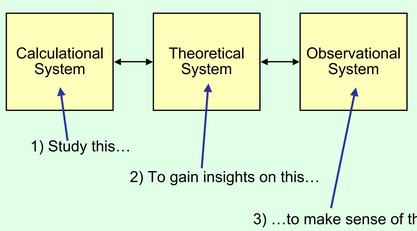
- the process by which models are selected and configured for a given scientific question.
- the process by which model outputs are selected, aggregated and interpreted by a community of expertise in climatology.
- the software fidelity of the models (i.e. whether the running code is actually doing what the modellers think it's doing).
- the (often convoluted) history that begat a given model, along with the modelling choices long embedded in the code.
- variability in the scientific maturity of different components within a coupled earth system model.

These omissions mean that quantitative approaches cannot assess whether a model produces the right results for the wrong reasons, or conversely, the wrong results for the right reasons (where, say the observational data is problematic, or the model is configured to be unlike the earth system for a specific reason).

Furthermore, quantitative skill scores only assess specific versions of models, configured for specific ensembles of runs; they cannot reliably make any statements about other configurations built from the same code.

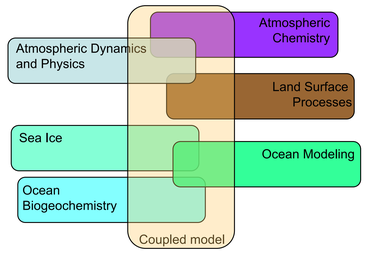
Quality as Fitness for Purpose

Evaluation of climate models should not be about "the model", but about the relationship between a modelling system and the purposes to which it is put. More precisely, it's about the relationship between *particular ways of building and configuring models* and the *ways in which the runs produced by those models are used*.

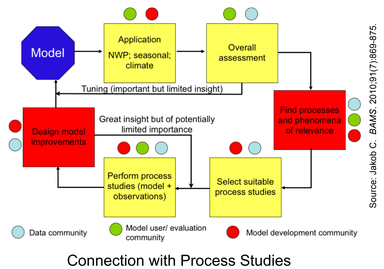


What are the uses of a climate model?

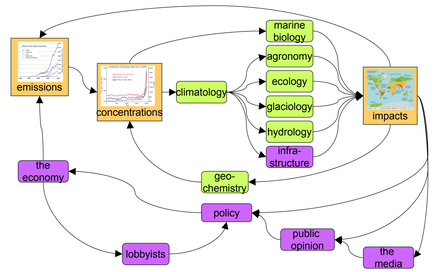
- To provide inputs to assessments of the current state of climate science;
- To explore the consequences of a current theory;
- To test a hypothesis about the observational system (e.g. forward modeling);
- To test a hypothesis about the calculational system (e.g. to explore known weaknesses);
- To provide homogenized datasets (e.g. re-analysis);
- To conduct thought experiments about different climates;
- To act as a comparator when debugging another model;



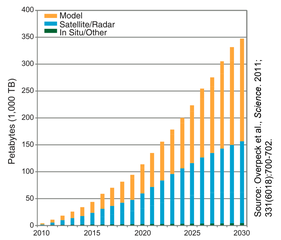
Scientific priorities, Communities of expertise



Connection with Process Studies



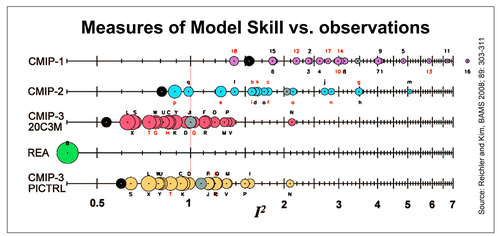
Use by broader set of stakeholders, Development of climate services

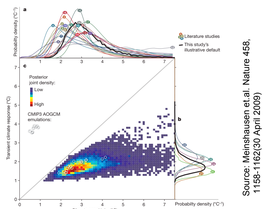


Working with data (re-analysis, homogenization, downscaling..)

Scientific and Societal Context

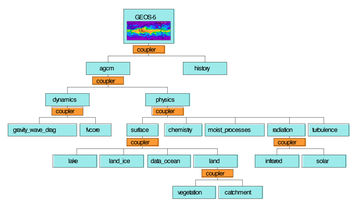
connect to..



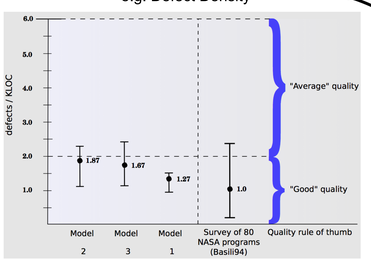


Science Communication Presentation of projections, Treatment of uncertainty

Use of software frameworks standards, interface definitions



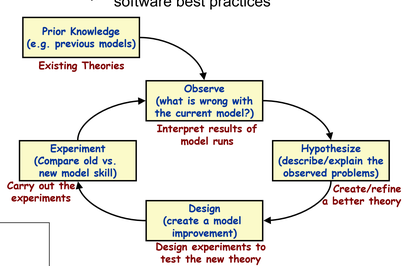
Software quality metrics e.g. Defect Density



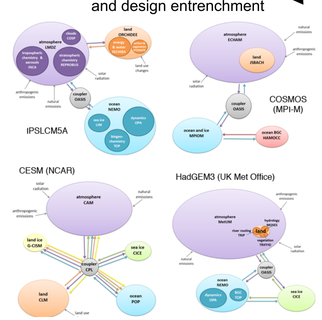
Model Development Processes

connect to…

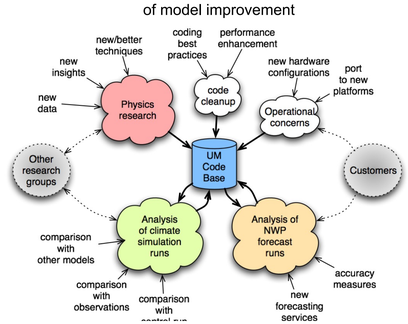
Comparison with software best practices



Software architectural choices and design entrenchment



Day-to-day process of model improvement



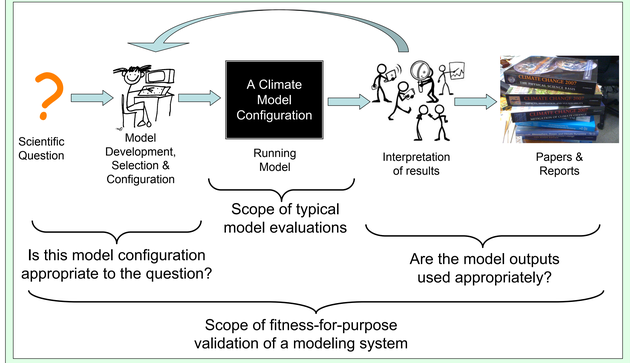
Testing Practices and test Coverage

Simplified tests (e.g 3D -> 2D)
Single-column tests
Dynamical core tests
Idealized aqua-planet test
Uncoupled model tests
Double-call tests for new versions
Spin-up tests
Conservation tests
Bit Reproducibility tests
Robustness tests (e.g many compilers)

Validation of the entire modeling system

When we ask questions about likely future climate change, we don't ask the question of the calculational system, we ask it of the theoretical system; the models are just a convenient way of probing the theory to provide answers.

When society asks climate scientists for future projections, the question is directed at climate ***scientists***, not their models. Modellers apply their judgment to select appropriate versions & configurations of the models to use, set up the runs, and interpret the results in the light of what is known about the models' strengths and weaknesses and about any gaps between the computational models and the current theoretical understanding. And they add all sorts of caveats to the conclusions they draw from the model runs when they present their results.



Validation is ***not*** a post-hoc process to be applied to an individual "finished" model, to ensure it meets some criteria for fidelity to the real world. In reality, there is no such thing as a finished model, just many different snapshots of a large set of model configurations, steadily evolving as the science progresses. Knowing something about the fidelity of a given model configuration to the real world is useful, but not sufficient to address fitness for purpose. For this, we have to assess the extent to which climate models match our current theories, and the extent to which the process of improving the models keeps up with theoretical advances.

Conclusions

Our approach to model validation extends current approaches:

- down into the detailed codebase to explore the processes by which the code is built and tested;
- up into the broader scientific context in which models are selected and used to explore theories and test hypotheses.

By taking software testing into account, we build up a picture of the day-to-day practices by which modellers make small changes to the model and test the effect of such changes (both in isolated sections of code, and on the climatology of a full model). The extent to which these practices improve the confidence and understanding of the model depends on how systematically this testing process is applied, and how many of the broad range of possible types of testing are applied. We also look beyond testing to other software practices that improve trust in the code, including automated checking for conservation of mass across the coupled system, and various approaches to spin-up and restart testing.

By taking the broader scientific context into account, we examine how features of the entire scientific enterprise improve (or impede) model validity, from the collection of observational data, creation of theories, use of these theories to develop models, choices for which model and which model configuration to use, choices for how to set up the runs, and interpretation of the results. We also look at how model inter-comparison projects provide a de facto benchmarking process, leading in turn to exchanges of ideas between modelling labs, and hence advances in the scientific maturity of the models.

This layered approach does not attempt to quantify model validity, but it can provide a systematic account of how the detailed practices involved in the development and use of climate models contribute to the quality of modelling systems and the scientific enterprise that they support. By making the relationships between these practices and model quality more explicit, we expect to identify specific strengths and weaknesses the modelling systems, particularly with respect to structural uncertainty in the models, and better characterize the "unknown unknowns".