Improving the Quantitative Interpretation of Simulation Models

Banff International Research Station for Mathematical
Innovation and Discovery (BIRS)
March 13 - March 18, 2016
Rapporteurs: R. Rosner & L. Smith

I. Introduction

Our workshop took place as planned, involving ultimately a group of 13 scientists, ranging from modelers to the "consumers" of models. Our special focus was on the modeling of terrestrial climate, from the relatively short-term (viz., weather to seasonal forecasting) to the long-term (e.g., decadal and climate forecasting); and our particular emphasis was on reaching a better understanding of what current models of terrestrial climate are capable of doing, how their present capabilities match with the needs of the "climate model consumers", and what will need to be done to accomplish a more satisfactory match between what can be done – even in principle – and what is desirable from the public policy perspectives.

The participants of our workshop are listed in the table immediately below.

Name	Affiliation
Berger, Jim	Duke University
Bogdan, Tom	N/A
Du, Hailiang	Univ. of Chicago (Computation Institute)
Mason, Simon	The Earth Institute of Columbia University
Nissan, Hannah	Columbia University
Oberkampf, William	WLO Consulting
Petersen, Arthur	University College London (UCL)
Rosner, Robert	University of Chicago (Dept. of Physics and Energy Policy Institute at Chicago)
Smith, Leonard	London School of Economics (and Pembroke College Oxford)
Stainforth, Dave	London School of Economics and Political Science
Tribbia, Joe	National Center for Atmospheric Research
von Hardenberg, Jost	Institute of Atmospheric Sciences and Climate – National Research Council
Wehner, Michael	DOE Lawrence Berkeley Laboratory-Scientific Computing Group

The discussions focused on two distinct areas of climate modeling: first, gaining a better understanding of model weaknesses; second, identifying specific areas that can lead to enhanced predictive capabilities of climate models. We provide a short glossary of key terms in an appendix.

II. Diagnostics of model weaknesses

1. Assessing model failure

The question of how to assess model failure entails a number of separate issues that are best discussed individually.

- *Current practice is to run models to provide predictions for the end of the Century*. This is not scientifically sensible: It would make scientific sense to run the forecasts out as far as the models can produce some meaningful information. Thus, model forecasts may have drifted off on highly unrealistic trajectories as a result of structural model errors after a relatively short period (only a few months or years); it makes no sense to apply simplistic bias corrections to these forecasts.

- *Quantifying the timescales on which different climatic variables are captured by an ensemble is a key desirable*. The above begs the question of how one is to discover how far out forecasts can be run before they stop being sensible. One possible solution is to look at model ensembles, which can provide a collection of trajectories of the future. The key is to identify the drivers, and to get them correct – this task is thus about assessing the timescale (Tau) on which the range of the ensemble fails to capture behavior in reality, and in particular the behavior of the drivers (example: El Nino). This could be as a function of some independent variable, for example, location **x**, e.g., Tau(**x**). Note that this does require us to distinguish between driving phenomena and target phenomena. For example, not being able to forecast a target phenomenon at time t^* need not imply we cannot forecast it at $2t^*$, but rather reveals our inability to forecast **the drivers of** the target phenomena. Multi-model (MM) ensembles can help us understand these connections, and the model properties that enable them, e.g., fidelity of the driver(s).

    - Example, in long-term simulations: When does bias correction (or projection of results from model space to reality) start to fail? Even if we bias-correct the models locally so that their climatology locally looks right, this will fail after some time. We could measure this, and define a local Tau(**x**).
    - Example in seasonal prediction: We typically initialize an ensemble of initial conditions with the observed state – and then we can ask, when does some model solution(s)/ensemble diverge clearly from reality (unable to shadow)? This would define Tau(**x**).

- *An important issue is the ability to identify where/when model failures occur today, e.g., using today's models*. One way to proceed is to identify geographic locations (viz., **x** = S. Africa/strong El Ninos; El Nino 2014) where Tau(**x**) is relatively short, e.g., where seasonal forecasts are known to fail.

Identifying regional failures, i.e. times and locations where a climate prediction ensemble fails in the sense that the model is unable to produce any ensemble members that represent the observed state, is an important example of what we aim for. Similarly, one could ask – alternatively, for multi-model ensembles (MMEs) – if some model members do get certain phenomena right, and others don't? In fact each model might outperform every other model robustly on some set of phenomena (in the short term), given the number of targets: Perhaps that can be used to identify the physical mechanisms for the model failures? This affords the opportunity for case studies on short-term climate that examine individual members of each ensemble in detail. Tracking the subsequent temporal evolution of such "local in time and space" failures also permits the determination how the forecast loses utility at longer times and distant regions.

2. *Optimal use of multi-model ensembles*

An important issue is relates to better ways to use multi-model ensembles (MMEs) in the extrapolatory range. Thus, in-sample comparison with past measurements (which have been used to guide model development) is potentially misleading – unless, of course, the model fails in such comparisons.

- Consider a multi-model ensemble where each model contributes an initial condition (IC) ensemble. As time passes, each IC ensemble tends to separate into distinct future distributions. Does this separation indicate (a) different initializations (i.e., the models would shadow each other given slightly different initial conditions), or (b) different emphasis in the physics of each model, or (c) physical inconsistencies between the models, or (d) identifiable loss of fidelity in some subset of models?

  Surely one aim of a MME is to yield distributions of trajectories ("tubes") that disagree in the fine details but agree on the big picture. When a coherent "big picture" ceases to exist, do we have any confidence in any of the individual models (as we know the details matter in order one phenomena)?

- How do we interpret the results of a multi-model ensemble? What types of conclusions might we draw, other than subsets of models are "obviously erroneous" or "obviously consistent"?
  - When different MM ensembles offer very different outcome worlds, and we agree that the detail of implementation (things we know about) have a first order impact, we are faced with (at least) two options: (1) the outputs are at best mis-informative; or (2) we find a method of using "fuzzy" probabilities under the claim/hope that each ensemble under each model has some hope in Hades of saying something warm and fuzzy.

    - Note that on the time scales over which very different model structures yield similar future distributions, we then know that the (implementation of) details (we know of) do not matter.
- We suggest that we can identify erroneous (unreliable) predictions by examining a set of output quantities from each of the models to determine if any of the set of output quantities (from any given model) is physically unreasonable, and so obviously erroneous. (This would be an example of the use of expert judgment.) Since we have no data about the future, this could be used as a surrogate for observational data, and could be a possible way to quantify a value of Tau, i.e., a minimum length of time for which we do not have evidence that the model is unreliable.

3. Using dynamics to improve model trajectories.

When model trajectories are not matching reality, one can explore model inadequacy by finding ways to "nudge" the model trajectory, to be close to reality. The needed nudges (which may also be large, e.g., "bashes") might indicate features of the model that could be problematic.

As another possibility, one can use the existence of shadowing orbits to find better initialization procedures which in turn produce better distributions of trajectories. If this cannot be done, or if the suggested initializations are unreasonable, this indicates a limitation of the model.

We note that the notions of "shadowing" and "nudging" need to be clearly differentiated. A model can shadow if a trajectory exists that stays "close" to the target time series of states. (N.b.: there are three definitions of "close", leading to epsilon-shadows, iota-shadows and phi-shadows). A forecast system can fail to be informative even when the model can shadow, simply because, say, the ensemble generation scheme failed to locate initial conditions which shadow. Alternatively, there may be no trajectory of the model (with probability one) that shadows. In this case, there may be pseudo-orbits that stay close to the targets. (Pseudo-orbits are not trajectories of the model, but segments of trajectory that are periodically "nudged" (thus, violating the dynamics of the model) so as to stay on/near track.)

III. Enhancing Predictive Aspects of Climate Modeling

The question we address here is how one might go about identifying aspects of climate change that are both of interest to the "users" of models, and are plausibly forecast with some "skill" by the models.

1. *We identified a potentially productive approach: Define a set of events that historically have had very low probability, but whose probability is increased significantly as a result of climate change.*

The aim is to avoid *post hoc* attribution of events by recasting the prediction problem for climate change. This could be done by *a priori* identifying a set of events that would not be expected to be observed without climate change. The observation of a sufficient number of such events would therefore be evidence for climate change. What we want to do is to define a set of events taking into account variables for which we can detect significant change (and we expect significant change on a physical basis); furthermore, we will need to determine what that "sufficient number" ought to be. The key element here is that we define ahead of time what might change, instead of doing attribution of events that have been observed, after the fact.

All this will involve completing the following tasks:

a. Use our physics knowledge to define rare events that may be unprecedented, and might occur more frequently because of climate change. Physics will also help to identify what would be the mechanisms (teleconnections, changes in blocking etc., changes in ocean circulation)

b. Identify variables that are most useful to users and to the general public. We may wish to construct different baskets based on different user groups.

c. Use statistics to define how many samples would be needed in order to characterize significant change. Exploit the fact that one could operate at different spatial and temporal scales to get more samples. Exclude variables where this is not possible. Empirical sampling from model simulations can also help to assess statistical significance.

d. Construct proof-of-concept analyses/experiments using model simulations.

e. An additional idea is that for each basket we could build one single index of change (possibly similar to the idea of climate hot-spots indices). This could be tracked historically. See for example the HY-INT index of hydroclimatic change by Giorgi et al. 2011 (doi:10.1175/2011JCLI3979.1.)

The goal articulated above can also be addressed by examining changes in the distributions, such as changes in the tails of the distributions (e.g., changes in extreme precipitation), or changes in, viz., the position of quartile distributions. The idea would be to predict expected changes in distribution functions for some particular observable, based on our physics/model understanding of climate change, and then to ask whether historical comparison of distributions for such observables show significant differences. It will be important to identify such observables for which such comparisons have NOT already been made, since the strength of this approach is to make a priori predictions (not post-dictions) of expected climate change-driven distribution function evolution.

2. "Packaging"

The idea is to make information usable for the greatest number of customers through a nested binary system. The core message is a binary one:

(a) I need to pay attention,

or

(b) Not to worry.

If (a), then there are two options:

(i) Our capacity to predict the future is severely compromised

and

(ii) We have reason for concern – at best, we may know that the chance of exceeding a user-defined threshold is greater than a certain level of interest.

This leads to a traditional stoplight ("tercile ") chart partitioning: Green (no worries) Yellow (pay attention: I have lost forecast capability), and Red (I have evidence that thresholds will likely be exceeded).

Events might include specific local events such as floods or intense precipitation, or might be regional impacts such as drought in South Africa or, more than x stations experiencing some climatic phenomena.

It is critical that the structure and flow of these decision trees be transparent, and may prove critical that information structures beyond those of probability be considered (non-probabilistic odds, for example).

3. *Studying and comparing ensembles from a range of existing models*

The idea here is to take advantage of existing (and possibly additionally computed) ensembles, based on existing models, to extract information that can inform risk assessments. For example:

- *Diagnose model errors*. Substantial differences in model results likely point to model inadequacies in at least one of the models. Diagnosing why the models diverge should provide useful insight into limitations of specific models and provide indications of what physical processes need to be modeled more accurately. Once these inadequacies have been identified, the relevant sub- models could be identified and modified to improve the sub-model's representation of the real world.

- *Identify model limitations*. These diagnostics of specific model weaknesses could be used as guidance to forecasters on the limits of a model's usability. For example, if a model tends to persist El Nino features too long into the boreal spring, we may decide to only use this model for seasonal forecasts out to late summer.

- *Inform forecasters*. How can we use our understanding of the observed climate dynamics of a region, combined with a diagnostics of the model climate to produce forecasts? For example, imagine we are trying to make a prediction at climate change timescales for South Asia. We would start with identifying the important controls on climate variability in the real world, such as the monsoon trough. We would then examine how the monsoon trough changes in the model climate. This approach would be more useful than downscaling, especially in areas with minimal global model accuracy.

4. *Will there be "changes" in predictability, in the future, on short time scales (say days to weeks) due to changes in the climate system over long time scales (decades to centuries)?*

To answer this question, we're led to ask the following:

a. How would we design experiments to shed light on this question?

b. What are the diagnostics (both observational and ensemble-based) we would use in order to answer this question?

c. Which (if any) changes would have the greatest consequence? Be easiest to address?

IV. Final Discussion

Our final discussion led us to identify four distinct activities that appear to be feasible at this point in time:

1. In order to progress, it is evident that we'll have to show progress on at least one seasonal test case – such as, for example, the El Nino/S. Africa connection. The key will then be to see whether one can actually demonstrate that something useful can be learned.

2. Next, we believe that "floating" a test set (= "basket") of events that can be used to "detect" or "confirm" our understanding of climatic change is both possible and will prove useful. Issues will include how we chose such a test set, and how we manage to stay away from the bugaboo of "full attribution". This approach is somewhat risky because we are not in a position to evaluate the likelihood of event classes occurring in the future that have (on the basis of what we know about the past) extremely low probability of occurrence (viz., have never happened before). A similar concern (without the risk) will plague attribution studies.

3. In order to deal with the potential limitation of #2, we might "float" a set of event distributions which we have reason to believe (on the basis of, e.g., expert

judgment) are likely (no computed probability!) to change significantly over some period into the future – with an aim similar to that of #2. The choice of event distributions should stay away from event distributions for which such comparisons have already been made – we want to do the choice a priori …

4. Position papers. We identified four domains in which it will be useful to "plant our flag":

    a. The solicitation and presentation of expert judgment and potential use of "fuzzy" probabilities, non-probabilistic odds, and so on.
    b. The development of rigorous techniques for capturing expert judgment

    c. Provide examples of traceable accountings of uncertainties for every significant example in the Summary for Policy Makers (SPM), as suggested in the current guidance notes..

    d. When do you pull the plug on a model, i.e., what are the diagnostics that tell us that models are becoming inadequate (n.b.: "inadequate" could refer to general inadequacy, i.e., a model is basically useless for any interesting forecasting, or could refer to inadequacy in a particular forecasting domain). Perhaps we can also discuss diagnostics that tell us that a model is adequate – probably a much harder problem. Finally, discussions of model failures may well be very informative for further model development and improvement, and improved physics understanding.

Appendix: Vocabulary

Bias: In the context of seasonal prediction, bias is the temporally varying (with month of the year) component of the systematic error(s).

Ensemble: A set of model simulations (which may be differentiated by different initial conditions, different underlying models, or different model parameters)

MME: A Multi-Model Ensemble is a set of simulations produced using different climate models

Nudges: Small perturbations to a model trajectory designed to accomplish some particular aim (keeping the trajectory near a sequence of observations, for example).

Shadow trajectory: A model trajectory that remains near (shadows) a sequence of observations to within a given tolerance. Iota-shadows, for instance, remain close enough to the target observations that one could argue the observations were in fact generated by the trajectory, given the observational noise model. The existence of a shadowing trajectory does not guarantee it would ever be found in an operational

ensemble, but the time scales on which no trajectory can shadow reveals a true limit of predictability of that model..