Defining expectations: an approach to quantifying trust in modelling

G. Abramowitz[a]

[a] *ARC Centre of Excellence for Climate System Science and Climate Change Research Centre, UNSW*

Email: gabriel@unsw.edu.au

Abstract: Developing 'trust' in a model involves experience with its predictive capacity in a range of environments, and in particular, the ability to identify circumstances in which we can quantify expectations of performance. In this talk I'll argue that an inability to clearly define our expectations of model performance has led to model evaluation frameworks that are essentially qualitative, and, most importantly, fail to answer the fundamental question of whether or not we have a 'good' model. While this aim might sound subjective, there are a number of ways one might prescribe threshold levels of performance *a priori* – before running a model – that could help define a 'good' model. These include:

1. *Better than another model*. In some subset of regions, variables and metrics we are interested in, a model outperforms another candidate model. While this is the benchmark we are all probably most familiar with, it is potentially a very weak benchmark. It might guarantee incremental improvement of successive model generations, but it does not discount the possibility that both candidate models are very poor.
2. *Fit for a particular application*. An example of this might be the ability to achieve threshold rates of moisture recycling that enable adequate representation of monsoonal phenomena. Once we meet such a threshold, we know that a model is appropriate for its intended use. Ideally, we would have clear benchmarks of this nature for hydrological, ecological, climate and weather applications as a minimum standard. Defining these thresholds meaningfully is in practice extremely difficult, particularly in coupled modelling systems.
3. *Utilisation of information*. This approach defines benchmark levels of performance as a function of the complexity of the model and the amount of information provided to it in its inputs and parameters that is relevant to the quantities it is required to predict. For example, a hydrological or land surface model that is given distributed soil and vegetation information in addition to meteorological forcing should be expected to perform better than one that is not. The same should be true of a non-linear model as opposed to a linear model.

I'll outline an example where this third approach was successfully used to illustrate that a collection of 13 international land surface models were significantly underutilising the available information in their meteorological forcing data about sensible and latent heat fluxes. Despite significant data uncertainties, it was found that an out-of-sample linear regression outperformed all land surface models' prediction of sensible heat flux across 20 flux tower sites globally using four different statistical metrics. While this type of exercise requires a significant volume of process level observational data at the scale of a model's application, this example has opened up a range of difficult questions for the land surface modelling community that are likely shared by other communities were appropriate observational data available or similarly applied. These include:

- What does it mean to say we have "physically-based" model of a natural system when we don't have enough data to construct an empirically-based model? How do we know our conceptual representations have any value in the absence of observations that can directly confirm process representation?
- Is the drive to add more processes (often based on very sparse data sets) leading to intractable modelling systems with relatively poor accuracy?
- Are inappropriate values for the unconstrained parameters (through calibration) actively inhibiting predictive ability at global and regional scales?

These and more are offered as possible discussion material for the session.

Keywords: *Model benchmarking, land surface models*