Journal of Statistical Planning and Inference 139 (2009) 1243 – 1245

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Discussion of reified Bayesian modelling and inference for physical systems by Michael Goldstein and Jonathan Rougier

Michael Lavine[a,*], Gabriele C. Hegerl[b], Susan Lozier[c]

[a] *University of Massachusetts, Amherst, MA 01003-9305, USA*
[b] *School of Geosciences, The University of Edinburgh, UK*
[c] *Duke University, Durham, NC 27708-0251, USA*

We congratulate Goldstein and Rougier (GR) for a fine paper on the difficult and important subject of what can be learned from imperfect computer simulators of physical systems. That this assessment is difficult is widely recognized. The contribution of GR is to propose a fully coherent framework for considering multiple simulators simultaneously. Our discussion is tripartite. First, we raise a point of clarification; second, we compare GR to current practice among climate modellers; and third, we discuss how GR can run into practical problems.

Clarification: GR's formulation relies on a reified simulator f^* and an input vector (x^*, w^*) as described by Eq. (8). In this formulation, (x^*, w^*) are system values. I.e., they are the true state of the real system we are trying to simulate. Our question is: "*Do such values exist*?"

In GR's example of the thermohaline circulation (THC), their reified simulator does not contain any further inputs or regressor functions than their generalized simulator f' (Eq. (29)). Thus f^* is a function of seven inputs: $(T_2^*, T_1^* - T_2^*, T_3^* - T_1^*, \Gamma, K, q, T_5^*)$; see Table 1.

GR's system is the Atlantic, though for simplicity they sometimes think of their system as the computer program CLIMBER-2. With respect to these systems, T_2^*, for example, is the temperature forcing of the North Atlantic. But neither the ocean nor CLIMBER-2 has a single T_2^* because neither system has a single atmospheric temperature over the entire North Atlantic. Temperature varies with location. Even in the simplified world of CLIMBER-2, temperature varies with latitude. Similar comments apply to other inputs and to other systems that are more complicated than their reified simulators. As a general rule, systems and their inputs are more complex than their simulators. So we ask GR to clarify what is meant by (x^*, w^*) both in general and in the example of the THC.

Contrast with current practice: In discussing GR, a useful point of comparison will be current practice among climate modellers for assessing uncertainty. We begin by noting that climate modellers are aware of the desirability of constructing distributions for uncertain parameters. A summary of some of their recent efforts can be found in Manning et al. (2004). There are several techniques in use; here we describe one used by Murphy et al. (2004) for the quantity called climate sensitivity which is, by definition, the equilibrium change in globally averaged surface temperature due to a doubling of atmospheric CO_2. We use the symbol α for climate sensitivity.

One result reported by Murphy is "a probability density function for the sensitivity of climate to a doubling of atmospheric carbon dioxide levels." Murphy et al. obtain the probability density function (pdf) by sampling tuneable parameter values uniformly over a range suggested by expert knowledge, running a climate simulator once for each draw of the parameters, weighting the simulations by how well they match current climate, extracting α from each simulation, and then assembling the weighted draws of α into a posterior density.

It is interesting to see how differently Murphy and GR use simulator runs. Murphy's simulator runs go directly into the posterior distribution. GR's simulator runs go indirectly into the posterior; they go into improving the emulator for f by updating

[*] Corresponding author. Tel.: +1 413 5450560.
E-mail address: lavine@math.umass.edu (M. Lavine).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2008.08.011

the posterior distributions for the quantities in the following equation:

$$f_i(x) = \sum_j \beta_{ij} g_{ij}(x) + u_i(x) \tag{9}$$

which in turn contribute to the posterior for α through the link from f to f^*. To see one practical difference between direct and indirect use of simulator runs, consider what happens to large simulated values of α. Murphy et al. put those values directly into their posterior. The posterior range of plausible α's increases, at least when the simulator run has a reasonably large weight. GR, in contrast, use those values of α to update the distribution of the β_{ij}'s and the u_i's. If a simulator run with a large α also has a large leverage, the result may be to move the distributions of both the β_{ij}'s and the u_i's closer to 0, possibly resulting in a smaller posterior range for α. It is worth considering whether this possibly unforeseen consequence is desirable.

On the other hand, Murphy's approach necessarily results in a posterior range for α not greater than the range observed in simulator runs. Most climate modellers recognize that the appropriate range of uncertainty should be larger than that observed in simulator runs, to account for several sources of variability including that associated with the choice of one particular simulator. GR have the potential to produce larger ranges for α, as do some of the other methods currently in use by climate modellers. See the IPCC report by Manning et al. (2004) for a brief discussion.

Nonetheless, the use of ranges larger than observed is not universally accepted. As noted by another recent report from climate modellers (Stainforth et al., 2005), large values of α "are not used in ranges for future climate change because they have not [previously] been seen in general circulation models." The contribution of Stainforth is to report the result of simulator runs that expand the range of α's previously simulated. They find large ranges by running many more simulations than previously possible using *climateprediction.net*, a distributed computing project to which the general public may donate unused cycles on their personal computers. Again, a point to note is that simulator runs are used directly in assessing the uncertainty for α.

Another point of contrast between GR and Murphy is that Murphy et al. are content to leave their pdf "contingent upon the structural choices made in building our GCM, the use of a linear prediction scheme, the choice and application of observational constraints and the choice of parameters for perturbation" and not to quantify the extent to which those choices may influence their pdf. GR, on the other hand, try to do away with the contingency by actively assessing the influence of those choices, including the choice of simulator, on the posterior. Clearly this is an important problem: two recent estimates of climate sensitivity based on the ability of models to simulate tropical sea surface temperature changes in the Last Glacial maximum (Schneider von Deimling et al., 2006; Annan et al., 2005) are quite different—the best-fit value of one falls outside the central 90% interval of the other—and a substantial part of that difference may be attributable to structural differences in the climate models used. If successful, accounting for structural uncertainty in the simulator would be a major step forward. GR's accounting is a combination of expert opinion and calculations with different priors.

Practical limitations: In principle it is advantageous, or at least not disadvantageous, to work in a coherent framework. In practice, however, there are several practical considerations which may limit the advantages of the coherent framework.

Hard work: Implementing GR takes a lot of intellectual work. First, one must construct an emulator for the currently available simulator f. That means working from GR's Eq. (9). Ideally, one would use physical insight to choose the basis functions g. Often, however, physical systems and their simulators are sufficiently complex that even experts have little insight into what collection of basis functions would adequately capture system behavior. That would be the case, for example, in climate simulators. So in practice one might simply choose linear functions of each input separately, as GR have done.

Also, one would hope for at least a minor amount of insight to model u, the lack-of-fit process. For us—a statistician, a statistical climatologist, and an oceanographer—insight into u is lacking.

But the emulator for f is probably the least important part of the modelling exercise because data, i.e., model runs, will update our possibly subjective priors into what we hope will be posteriors that can be widely accepted. So let us not spend too much time on Eq. (9) and pass, instead, to more crucial aspects of modelling.

The next modelling step is to

> "consider a version of the simulator with enlarged set of inputs, $f^*(x, w)$ say, which is sufficiently careful ...that we would not consider it necessary to make judgements about any further improvement …" (pg 4)

One problem here is that we are not sure what this passage means. In GR's example is f^* supposed to be an ocean simulator with hugely many compartments? In our view, an f^* for the THC would have hundreds, if not thousands, of compartments. Thus the set of inputs would be enlarged by several orders of magnitude.

Is f^* supposed to be so accurate that its differences from the actual physical system are negligible? If not—if f^* has nonnegligible errors—then why is it not necessary to contemplate further improvements?

But even supposing we can imagine f^*, we must also contemplate its relationship to f. That means working with GR's Eq. (10)

$$f_i^* = \sum_j \beta_{ij}^* g_{ij}(x) + \sum_j \theta_{ij}^* h_{ij}(x, w) + u_i^*(x, w) \tag{10}$$

and considering how the β^*'s are related to the β's. Here, as opposed to Eq. (9), the modelling and prior assessments are crucial because they do not get updated by data. Since f^* does not exist there are no model runs. If we get it wrong, there is nothing to save us. This part of the subjectively assessed prior does not get updated. Therefore we must make very careful prior assessments to get our uncertainties at least approximately right.

But getting it right is difficult. To bring home that point, we note that in their example, GR have simplified by "not introduc[ing] any further simulator inputs or regressor functions." To assess the value of GR we will, at some point, need to see examples without this simplification.

To compound the difficulty, we may need to work in high dimensional spaces. In particular, the dimensions of θ^* and w may be huge. Finding believable priors for these quantities will be difficult.

Technical compromises: Implementing GR requires high dimensional prior assessments and that those assessments be believable. GR note that u "is usually taken to be either a weakly stationary process or a stationary Gaussian process." These choices are convenient for both assessment and computational reasons, but they may not, in general, represent anyone's real prior. The point is that no matter how complicated a model we contemplate, technical compromises will be made for the sake of convenience, and these compromises may have unintended and unchecked influence on the end result.

As another technical compromise, GR work with a generalized simulator f' that requires only two more inputs than f, and with a reified simulator f^* that requires no more inputs than f'. Their prior for the quantities in Eq. (9) is probably not crucial because it is updated by model runs; their two-dimensional extension from f to f' is small enough to allow plausible prior assessment of the quantities in their Eqs. (26)–(28); and their nonextension from f' to f^* greatly simplifies their assessment and makes it believable by others. Also, they do a good job of checking sensitivity to prior assessments. But in order to judge GR fully we need to see how prior assessment and sensitivity checks will work with high dimensional extensions.

Scalability: We have already mentioned several aspects of GR that might cause practical implementation difficulties in high dimension. Current ocean-atmosphere climate models have on the order of 10^5 or 10^6 grid boxes, 10^6 inputs, and take on the order of months to run an ensemble simulation of 20th century and future climate on a supercomputer. Do GR consider these models f, f', or something else? *Climateprediction.net* uses a slightly coarser resolution, does not have a coupled ocean simulator, and using worldwide idling machines, has resulted only in tens of thousands of runs. Is it running f, f', or something else? Whatever the answer to these questions, it will be useful to see how GR handle a simulator of this scale.

Uncertain uncertainties: Climate modellers, policy makers, and the public would all prefer to have better estimates of uncertainty for, e.g., climate sensitivity or the imminent shutdown of the THC. Current estimates and their stated uncertainties are not taken at face value. GR can help if their uncertainties can be widely accepted. But can they? We believe the difficulties inherent in reifying f^* and assessing its connection to f are so great that estimates and their uncertainties will be accepted only with substantial reservations, even by those involved in the process. That is especially true for those parts of the assessment relying on subjective prior assessments that do not get updated. The qualms that GR evoke by referring to climate analyses that are "conditional on the model (or the simulator) being correct" (pg. 2) may be replaced by qualms about analyses being "for me personally, not necessarily for anyone else."

In our view, the outcome of any uncertainty analysis for computer simulators, including GR's, is an approximation to an ideal analysis. If a fully coherent framework can make a better approximation, so much to the good. We have tried to point out practical limitations that might affect the accuracy of GR's approximation. If we felt that direct use of simulator runs, or some other method, resulted in a better approximation, then we would use it. We might ask whether it is incoherent and, if so, how. But if we judged the incoherency to be so slight as not to seriously degrade the approximation, we would accept it.

Conclusion: This discussion has pointed to some difficulties with GR's approach. Nonetheless, we believe GR have done a service by providing a fully coherent framework for modelling uncertainties inherent in computer simulators. Just by taking on the task they have proved themselves braver than we. All attempts to fully quantify uncertainty in climate change have limitations at present. We are therefore eager to see in a practical implementation how far the reified simulator can go to address them. We congratulate GR and offer our encouragement and best wishes as we await further development.

References

Annan, J.D., Hargreaves, J.C., Ohgaito, R., Abe-Ouchi, A., Emori, S., 2005. Efficiently constraining climate sensitivity with paleoclimate simulations. Scientific Online Letters on the Atmosphere 1, 181–184.

Manning, M., Petit, M., Easterling, D., Murphy, J., Pathwardhan, A., Rogner, H.H., Swart, R., Yohe, G., (Eds.), 2004. Describing scientific uncertainties in climate change to support analysis of risk and of options, Report of an IPCC workshop, 11–13 May 2004, Maynooth, Ireland, available from ⟨ipcc-wg1.ucar.edu/meeting/URW/product/URW_Report_v2.pdf⟩.

Murphy, J., Sexton, D., Barnett, D., Jones, G., Webb, M., Collins, M., Stainforth, D., 2004. Quntification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430, 768–772.

Schneider von Deimling, T., Held, H., Ganopolski, A., Rahmstorf, S., 2006. Climate sensitivity estimated from ensemble simulations of glacial climate. Climate Dynamics, 10.1007/s00382-006-0126-8.

Stainforth, D.A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D.J., Kettleborough, J.A., Knight, S., Martin, A., Murphy, J.M., Piani, C., Sexton, D., Smith, L.A., Spicer, R.A., Thorpe, A.J., Allen, M.R., 2005. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433, 403–406.