A Statistical Approach to System Inference Using Models

Jonathan Rougier[*]

Research Fellow

Department of Mathematical Sciences

University of Durham, UK

January 31, 2005

I trust we all agree that—subject to the usual caveats about the sociology of science—we build models to understand complex systems. So let our starting point be the question "How can I reduce my uncertainty about the physical system by evaluations of my simulator?" (see the comment below, "A little terminology", for my preference for 'simulator' rather than 'model' in this context).

In conferences and seminars we often we hear statements of the form "The simulator predicts . . . ". Likewise, in respectable journals we often observe the phrase ". . . conditional on the simulator being correct" (actually, the

[*] Department of Mathematical Sciences, University of Durham, South Road, Durham DH1 3LE, U.K. Tel: +44 (0)191 334 3111; e-mail: J.C.Rougier@durham.ac.uk. Funded by the Natural Environment Research Council, grant ref. NER/T/S/2002/00987 (RAPID Thematic Program) and the Tyndall Centre, grant ref. T2/13.

statement is more usually "... conditional on the *model* being correct", but on my interpretation of 'model' this is not right). This is not acceptable in cases where the behaviour of the system, if unchecked, has the potential to impose huge costs, and if the policies required to ameliorate the impact of the system are themselves hugely costly. This is certainly the case in climate change, the area in which I work, and it is also the case in flood risk assessment. No expert should be allowed to get away with statements about the simulator when in fact the stakeholders require statements about the system.

The degree to which we expect our simulators to be informative about the system depends on our own prior assessment of simulator quality. In particular, when we have a sequence of simulators with the same underlying simulator structure but different resolutions of solver, we must have some mechanism for ensuring that the higher-resolution simulators are more informative in our inference about the system than the lower-resolution ones. But even where we only have one simulator, we should still be aware that our treatment of the information from that simulator must be coherent with the possibility that we could build a better simulator if we so chose. Therefore every time we use a simulator to understand a system we must ask the question "Where is the slot which allows me to quantify how good a simulator I believe this to be?"

Recent developments in statistics have tackled the problem of how we quantify the information about the system available in evaluations of a given simulator, an area where UK research groups can reasonably claim to be world-leaders. From a statistical point of view we identify two sources of uncertainty. First, the simulator contains parameters about which we are uncertain. Sometimes these are measurable quantities that we have not been able to measure, like the initial value of the state vector in a dynamic system. More often, however, they are non-measurable in the context in which they enter the simulator. Hydraulic resistivity, for example, is difficult to measure at a point, and also difficult to generalise to a region treated as homogeneous. In ocean simulators it is necessary to distinguish between molecular viscosity, the measurable quantity in the model and the system, and eddy viscosity, the number that goes into the simulator, which is typically several orders of magnitude bigger.

Second, even if we were able to identify, perhaps through some supernatural agency, the best value for these unknown parameters, then we do not believe that the simulator evaluated at those parameters will exactly match the system. Note that this is a much stronger requirement than exactly matching the data, or, indeed, exactly matching the system values that correspond to the data (i.e. without measurement errors). This gives us two uncertain quantities to specify: the 'best' parameter value x^*, and the *simulator discrepancy* ϵ. The simplest possible way in which we can combine these together with the simulator f and the system y is in the form

$$y = f(x^*) + \epsilon \qquad x^* \perp\!\!\!\perp \epsilon$$

where '$\perp\!\!\!\perp$' denotes probabilistically independent in the mind of the expert performing the analysis. This requires us to specify a distribution function for x^* and a distribution function for ϵ. This approach has been adopted by our group at Durham in a series of papers (see, e.g., Craig et al., 2001), in broad collaboration with groups at Sheffield (Kennedy and O'Hagan, 2001) and Los Alamos (Higdon et al., 2005).

This simple statistical framework is easy to interpret. It states that were x^* revealed to us, then for the purposes of making a prediction about y, we would be satisfied with a single evaluation of the simulator at x^*. Because x^* is not revealed to us in practice, we have to perform many evaluations of our simulator at candidate values for x^*, and we have to weight the outcome of these evaluations according to our prior beliefs about x^*. If we are lucky enough to have observations, z say, on y (albeit incomplete and made with error), then we weight by the conditional distribution of $x^* \mid z$. It is important to realise, however, that to compute this conditional distribution we must first specify the distribution of the discrepancy ϵ. It is not the case that having data for calibration can excuse us from thinking about the quality of our simulator. The point is, the better is our simulator, the more informative about x^* we expect the data z to be.

The simple approach with ϵ additive and independent of x^* is very attractive from a computational point of view, which is important given the general complexity of the inferential calculations. I would also say that it is quite intuitive. Certainly, it is possible to 're-interpret' a common and ostensibly non-probabilistic approach in these terms; see, for example, the experiment outlined in Murphy et al. (2004), which I re-interpret in Rougier (2004). It also leads to quite simple measures of simulator inadequacy, based on summary statistics such as $\mathsf{Var}(\epsilon)$. If you want to make probabilistic statements about the system, and you do not have a slot for simulator quality in your current set-up, then you would be well-advised to start with this approach. Even so, the specification of $\mathsf{Var}(\epsilon)$ is not straightforward, particularly when accounting for spatial and temporal indexing in y; Craig et al. (2001, section 6.1) provide an example of how we might proceed in this case. But what is the best thing to do? Make an honest attempt at appraising a difficult quantity, or simply ignore it (effectively set $\mathsf{E}(\epsilon) = \mathbf{0}$ and $\mathsf{Var}(\epsilon) = \mathbf{0}$). Practically speaking, the answer to this question depends on the requirements of the stakeholders. I believe that we should educate stakeholders to the point where they will demand quantifications of simulator inadequacy as a necessary prerequisite for simulator-based inference about systems.

Finally, a brief look-ahead. Although it may be considered the current state-of-the-art, the simple additive discrepancy term is too highly constrained to be an adequate reflection on the very complicated issue of simulator quality across a range of simulators for the same underlying system, which might vary in the ways described in the Note on Terminology. The 'next big thing' in this area will be more general formulations that allow us to link simulators together into a joint inference about the system. In climate prediction, for example, we need a way of making sense of seemingly inconsistent simulator-based assessments of quantities such as climate sensitivity (Stainforth et al., 2005, is the latest such estimate). In a recent paper (Goldstein and Rougier, 2005) we have outlined an approach for this, and we are currently involved in a case-study to examine practical implementation issues. It's a difficult job, but we should take as a guiding principle that we allocate effort according to the importance of the task. In climate, and perhaps in flood risk assessment as well, the simulators are rather poor and the consequences of unchecked system behaviour are potentially catastrophic, and so we should expect to devote a large amount of effort to quantifying the inferential content of simulator evaluations, including notions of simulator quality, as a necessarly prequel to making simulator-based predictions for system behaviour.

A note on terminology. I make a careful distinction between a model, its treatment and its simulator. Broadly, we may think of the simulator, which is the code we actually evaluate, as arising from

$$\mathsf{Simulator} = \mathsf{Model} + \mathsf{Treatment} + \mathsf{Solver}.$$

The model tends to be the underlying mathematical equations, often written as a collection of differential equations and equations of state. The treatment typically concerns the initial and boundary conditions (including forcing functions) that make the model applicable to a particular time and place. The treatment can also concern which properties of the model are taken as outputs: e.g., steady state, 'ergodic' averaging, or dynamic evolution subject to specified forcing. Finally, the solver requires decisions about discretisation, in particular the order of the approximation and spatial and temporal resolution.

None of these distinctions would be necessary if there was a one-to-one correspondence between models and systems. But when formulating a coherent framework linking models and systems it is essential to acknowledge that there are many *simulators* for a given system, and that these simulators share common features due to having similar models, treatments and solvers. We cannot think about how our simulator is informative about the system without also being prepared to think about how our simulator links to other simulators of the same system. For example, this is how we will be able to use palæo-climate data to calibrate future climate predictions. The palæo-climate-simulator and the future-climate-simulator share the same underlying climate model, and are linked up through uncertainty about the model parameters.

References

Craig, P., M. Goldstein, J. Rougier, and A. Seheult: 2001, 'Bayesian Forecasting for Complex Systems Using Computer Simulators'. *Journal of the American Statistical Association* **96**, 717–729.

Goldstein, M. and J. Rougier: 2005, 'Probabilistic Formulations for Transferring Inferences from Mathematical Models to Physical Systems'. *SIAM Journal on Scientific Computing* **26**(2), 467–487.

Higdon, D., M. Kennedy, J. Cavendish, J. Cafeo, and R. D. Ryne: 2005, 'Combining Field Data and Computer Simulations for Calibration and Prediction'. *SIAM Journal on Scientific Computing* **26**(2), 448–466.

Kennedy, M. and A. O'Hagan: 2001, 'Bayesian Calibration of Computer Models'. *Journal of the Royal Statistical Society, Series B* **63**, 425–464. With discussion.

Murphy, J. M., D. M. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth: 2004, 'Quantification of Modelling Uncertainties in a Large Ensemble of Climate Change Simulations'. *Nature* **430**, 768–772.

Rougier, J.: 2004, ‘Brief Comment Arising re: “Quantification of Modelling Uncertainties in a Large Ensemble of Climate Change Simulations” by Murphy et al (Nature, 2004)’. Unpublished, available at `http://www.maths.dur.ac.uk/stats/people/jcr/newMurph.pdf`.

Stainforth, D., T. Aina, C. Christensen, M. Collins, N. Faull, D. Frame, J. Kettleborough, S. Knight, A. Martin, J. M. Murphy, C. Piani, D. Sexton, L. A. Smith, R. Spicer, A. Thorpe, and M. Allen: 2005, ‘Uncertainty in Predictions of the Climate Response to Rising Levels of Greenhouse Gases’. *Nature* **433**, 403–406.