

# Statistical Approach: First Thoughts

Jonathan Rougier, Neil Edwards, and David Cameron

January 26, 2006

## 1 Background

The origins of statistics are in experimental design (e.g., agricultural field trials), and this is still a large part of contemporary statistics. A rapidly-developing field, referred to as *Computer Experiments*, is concerned with experiments that do not require physical implementation, such as those that take place using computer models (see, e.g., Sacks et al., 1989; Currin et al., 1991; Koehler and Owen, 1996; Santner et al., 2003). The critical feature of this field is the absence of unspecified sources of variation. Much of traditional experimental design is about performing replications to make sure that these sources of variation do not interact systematically with the response (see, e.g., Cox, 1958, a ‘classic’: not technical and well worth reading). Computer-based climate models are deterministic: they have no unspecified sources of variation, and consequently there is no need for replication.

The field of Computer Experiments belongs within the general area of *Model-Based Inference for Complex Systems (M-BICS)*. In our experience, the challenges of a particular M-BICS application can be broadly classified by (i) size of model; (ii) model inadequacy; and (iii) policy impact. Climate prediction scores

highly on all three. Rougier (2005) provides a general description of probabilistic ensemble-based inference for climate, i.e., current practice as interpreted (and critiqued) by a statistician.

Our objective is to bring statistical insights from recent developments in M-BICS to bear on the Modelling Intercomparison Project. *The principle objective is to create a statistical framework within which we can relate a hierarchy of climate models.* A simple use for such a framework would be to predict the behaviour of a particular model at a particular model-input on the basis of evaluations of that model, and also evaluations of other models. This framework is interesting in itself. For example, it allows us to summarise the basic structure of each model, or to quantify the ‘closeness’ of two models in terms of the ability of evaluations of one model to reduce our uncertainty about the other. It also helps us to design more informative ensemble experiments, for example by targeting evaluations of a particular model at model-inputs that are judged, *a priori*, to be good at reducing uncertainty, either about the model itself, or about a model further up the hierarchy.

At the very top of the hierarchy we have reality itself: the climate system. The *reified analysis* approach developed at Durham (Goldstein and Rougier, 2005a,b) describes a framework for linking the models to each other, and linking the very best model to climate itself. In this way evaluations on all models contribute to a climate prediction, although better models contribute more because they are ‘closer’ to climate.

## What is this ‘statistical framework’?

The starting-point is to appreciate that the climate model  $g(\cdot)$  represents an unknown function. It is unknown in the sense that  $g(x)$  is unknown unless we have actually evaluated the model at  $x$ . In our case  $x$  comprises model-inputs such as model parameters, initial conditions, and forcing functions, and  $g(x)$  are model-outputs (see below, section 2). For a given model our ensemble can be thought of as the collection  $(G; X)$ , where  $X \triangleq (X_1, \dots, X_n)$  are the model-inputs we have evaluated, and  $G \triangleq (g(X_1), \dots, g(X_n))$  are the resulting model-outputs. The ensemble represents an experiment to learn about  $g(\cdot)$ .

Using the ensemble, we can construct a statistical object called an *emulator* that allows us to predict  $g(x)$  on the basis of  $(G; X)$ . If  $x \in X$  then our prediction is exact. Otherwise our prediction involves some uncertainty, and this uncertainty tends to grow as the distance between  $x$  and  $X$  gets bigger. Emulators tend to be robust inside the convex hull of the ensemble (*interpolation*), but outside the convex hull (*extrapolation*) reliable emulators are rather more difficult to construct. One thing that catches some people out is the effect of dimension. A high-dimensional space is almost all corners, and so almost all of the model-input space is an extrapolation from the convex hull of  $X$ . Design for computer experiments is about good choices for  $X$ , so that the empty parts of the model-input space correspond to tractable responses in the model-outputs.

Our principle reason for constructing emulators is because they form the basis of our approach to linking models in a hierarchy. When we say that two models are similar, we mean that they tend to respond to similar model-inputs in similar ways. Our emulator takes the form of a mapping between the model-inputs and

the model-outputs. Similar models have similar mappings. For example, we might think of our emulator for model  $A$  as

$$g(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_1 x_2 + \dots$$

where  $\alpha_0, \alpha_1, \dots$  are uncertain coefficients that we learn about using our model  $A$  ensemble. Model  $B$  might have the same model-inputs but a higher resolution. The emulator for this model might be

$$g'(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \dots$$

The degree to which we judge that model  $A$  and model  $B$  are similar can be summarised in the way in which the  $\alpha$  and  $\beta$  coefficients are related. For example, in two very ‘close’ models the correlations between these two sets of coefficients would be almost 1. In this case the ensemble from model  $A$  would be highly informative about model  $B$ , in the sense that predictions for  $g'(x)$  based on the model  $A$  ensemble and the pair of emulators would have small uncertainties.

This framework can be generalised to models with nested model-inputs, and then to models with non-nested model-inputs (e.g., by embedding). The precise construction of the linkages across the emulators will depend crucially on the judgements of the modellers. Diagnostics can help us to identify mistaken judgements, but only expert insights can help us to find really good linked emulators. Note that ‘bad’ linked emulators are not useless, but they tend to have greater uncertainties going up the hierarchy of models.

## 2 Foreground

From a statistical point of view (and not *just* a statistical point of view!), the most important thing to establish at the very start of the project is an operational description of the Thermohaline Circulation (THC) itself. This is the ‘fixed point’ that allows us to relate the models to each other: each model’s outputs includes a representation (probably imperfect) of the same underlying climate quantities. We expect this will be  $O(10)$  quantities. If we want to calibrate the models using actual climate data, then these also need to be defined in operational terms, and included among the models’ outputs. We expect this will be  $O(100)$  quantities. Therefore each model’s outputs comprise two sets, informally ‘historical’ and ‘future’, and we can write  $g(x) \equiv (g_h(x), g_f(x))$ , where the dimension of each model’s output-space  $\mathcal{G} \triangleq \{g(x), x \in \mathcal{X}\}$  is  $O(100)$ . The dimension of the input-space  $\mathcal{X}$  will be model-specific. The GOLDSTEIN ocean model, for example, has 12 model-inputs.

For concreteness, we are focusing initially on the CO<sub>2</sub> ramping experiment, and on the climate models C-GOLDSTEIN, IGCM-GOLDSTEIN, FORTE, and possibly FAMOUS. In the proposal we favoured FAMOUS over FORTE, but we now judge, on the basis of model structure, that FORTE will be an important stepping-stone in linking the simpler models to FAMOUS.

Here are some interesting and relevant statistical issues:

1. These models share components: e.g., the GOLDSTEIN ocean, the IGCM atmosphere. Should the primitives in the statistical framework be the models, or the model-components?
2. What does it mean to say that model  $A$  is judged to be better than model  $B$  when the two models do not nest: is this a probabilistic sufficiency statement?

3. To what extent is it possible to learn about differences between models in a designed experiment (*variance learning*)?
4. How do we incorporate expert judgements into the choice of evaluations in an ensemble?
5. Can we use statistical methods (dimensional-reduction, emulation, stochastic optimisation) to shorten the time needed for spin-up?

The first two of these are foundational: we need to clarify them in order to construct an effective framework across the models (the groundwork has been laid in Goldstein and Rougier, 2005a,b). The third is largely experimental. The fourth and fifth are operational: they can help to make the experiment more efficient.

**One year deliverables.** Two factors constrain our one-year deliverables.

First, we want to make a contribution to the choice of evaluations in the ensemble, and this needs to be done at the start of the project.

Second, other projects will overlap with ours, and where possible we would like to take advantage of their progress. In particular, the *Managing Uncertainty in Complex Models* project is directly relevant (MUCM, PI Tony O'Hagan, <http://mucm.group.shef.ac.uk/>). This is a four-year RCUK-funded project with seven post-docs and four studentships, starting in June. In the US, the Statistical and Applied Mathematical Sciences Institute (SAMSI) are running a year-long programme in 2006-7, *Development, Assessment and Utilization of Complex Computer Models* (PI James Berger, <http://www.samsi.info/programs/>).

There is also the possibility of putting a strong PhD student onto the spin-up problem, starting in September.

On this basis we propose the following one-year statistical deliverables:

1. *Explore the possibility of installing FORTE at CEH-Edinburgh.*

As already explained, we see FORTE as a useful stepping-stone on the route to FAMOUS, and beyond. We would like to integrate FORTE with our ensemble design experiments on C-GOLDSTEIN and IGCM-GOLDSTEIN.

2. *Incorporate statistical insights into the designs of ensembles.*

This is not so much about choosing the precise evaluations in the ensembles, but rather introducing modellers to the general principles underlying experimental design, and widening the number of options available. Current designs in climate science tend to be single-parameter perturbations, full factorial, or simple monte carlo with a uniform weighting function. These are rather extreme designs. Better mainstream choices might include latin hypercube sampling, fractionated factorials, stratified sampling, importance sampling with variance reduction techniques (see, e.g., Robert and Casella, 1999; Evans and Swartz, 2000), or a more tightly-specified Bayesian approach (see, e.g., Chaloner and Verdinelli, 1995). These can be embedded within a sequential or batch-sequential approach for more efficiency (see Craig et al., 2001; Goldstein and Rougier, 2005c, for examples of sequential methods).

The right choice will depend on the precise objectives of the experiment, on expert judgements about the models, and on their hierarchical relationship. JR has experience of this from the ongoing QUMP experiment at the Hadley Centre, and the same ideas will be applied in the QUEST PalæoQUMP project (PI Sandy Harrison, starting in February).

3. *A framework linking C-GOLDSTEIN and IGCM-GOLDSTEIN.*

We choose a relatively simple starting-point for constructing a statistical framework, in the hope that some of the more difficult questions will be answered by other research projects during the course of the first year. Both of these models share the GOLDSTEIN ocean, and so this pair provides an application within which we can explore climate models that ‘overlap’. We focus initially on ‘future’ model-outputs, i.e., the  $O(10)$  description of the THC.

We aim to create ensembles of evaluations for both models, but to a design that is jointly-informative. We will also start to investigate experimental variance learning, with the input of Michael Goldstein (Durham, no relation). This work will complement the final year of JR’s funding on the NERC-RAPID project *The probability of rapid climate change* (PI Peter Challenor, NOC Southampton).

By the end of the year we should be able to demonstrate a framework in which we can make inferences about the behaviour of the IGCM-GOLDSTEIN model, informed by ensembles of evaluations from both IGCM-GOLDSTEIN and C-GOLDSTEIN. We ought to be able to answer questions like: what are the most important model-inputs, where are the non-linearities, which model-inputs show strong interactions? We ought also to have an approach for selecting new evaluations efficiently, addressing questions such as ‘should we do five evaluations of C-GOLDSTEIN or one evaluation of IGCM-GOLDSTEIN?’

### 3 Summary

The principle objective is to create a statistical framework within which we can relate a hierarchy of climate models:

- To summarise the response of any particular model to its model-inputs;
- To predict the behaviour of a particular model at a particular model-input on the basis of evaluations of that model, and also evaluations of other models;
- To allow us to quantify the similarity of two models in terms of the ability of one model to reduce our uncertainty about evaluations of another model further up the hierarchy;
- To link into climate itself—at the very top of the hierarchy—for calibration and prediction.
- *First year deliverables:* Installation of FORTE at CEH-Edinburgh (if possible), involvement in the selection of ensembles, and a framework linking the ‘future’ outputs of C-GOLDSTEIN and IGCM-GOLDSTEIN in the CO<sub>2</sub> ramping experiment.

### References

- K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.
- D. R. Cox. *Planning of Experiments*. New York: John Wiley & Sons, Inc., 1958.
- P.S. Craig, M. Goldstein, J.C. Rougier, and A.H. Seheult. Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, 96:717–729, 2001.

- C. Currin, T.J. Mitchell, M. Morris, and D. Ylvisaker. Bayesian prediction of deterministic functions, with application to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86:953–963, 1991.
- M. Evans and T. Swartz. *Approximating Integrals via Monte Carlo and Deterministic Methods*. Oxford: Oxford University Press, 2000.
- M. Goldstein and J.C. Rougier. Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM Journal on Scientific Computing*, 26(2):467–487, 2005a.
- M. Goldstein and J.C. Rougier. Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*, 2005b. Accepted as a discussion paper subject to revisions, available at <http://www.maths.dur.ac.uk/stats/people/jcr/Reify.pdf>.
- M. Goldstein and J.C. Rougier. Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association*, 2005c. Forthcoming, available at <http://www.maths.dur.ac.uk/stats/people/jcr/revision.pdf>.
- J.R. Koehler and A.B. Owen. Computer experiments. In S. Ghosh and C.R. Rao, editors, *Handbook of Statistics, 13: Design and Analysis of Experiments*, pages 261–308. North-Holland: Amsterdam, 1996.
- C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. New York: Springer, 1999.
- J.C. Rougier. Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, 2005. Forthcoming, pre-revision draft available at <http://www.maths.dur.ac.uk/stats/people/jcr/CCrevisionA4.pdf>, new version available shortly.
- J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989. With discussion, 423–435.

T.J. Santner, B.J. Williams, and W.I. Notz. *The Design and Analysis of Computer Experiments*. New York: Springer, 2003.