Bayesian Analysis (2008) **3**, Number 1, pp. 45–56

Comment on article by Sansó et al.

Jonathan Rougier*

1 Introduction

This paper represents a very welcome combination of Statistics and Climate Science. I am sure that no-one who has studied the paper is in any doubt about how demanding this type of collaboration is: it is splendid that statisticians and climate scientists are working together to understand better uncertainty in future climate.

As a statistician developing methods for computer experiments, I like climate science precisely because it is so challenging. In particular, the models are still quite poor on the scales for which we would like to use them (transient and regional behaviour). That is to say they have large *structural errors*: errors that cannot be removed simply by tuning the model parameters. They are also some of the most expensive models in the world to evaluate. Typical performance is about three model-years of output per day at the main research centres. Tony O'Hagan (2006) has termed the consequence of this paucity of evaluations 'code uncertainty'. In some applications we also have to contend with the scale of the model outputs: the state vector can easily have millions of components.

The MIT2DCM of Sansó *et al.* is of relatively low resolution, and in this case-study the focus is on just three uncertain model-inputs, so code uncertainty is not going to be a problem. As a consequence of the low resolution, though, and the small number of uncertain inputs, structural error is going to be crucial. For the last forty years, the trend in climate science has been towards higher and higher resolution models, and this will continue because so much of the important physics is missing even at current high resolutions (where a grid-cell in the solver is typically about 250 km a side). There are important questions concerning how much we can learn from low resolution models, and one of the projects I am working on addresses exactly that, by trying to understand the structural links between models along a spectrum of modelling refinements.

Sansó *et al.* are interested in calibrating MIT2DCM, i.e. using observations on climate to learn about the correct setting of this model's parameters. Probabilistic learning requires a statistical model that links (i) evaluations of MIT2DCM, (ii) the model's parameters, and (iii) observations on climate. A crucial component of this statistical model is the treatment of structural error. One thing I particularly like about this paper is that Sansó *et al.* have explicitly included a term for MIT2DCM's structural error (which they term ξ). Actually, in their treatment this term combines structural error, representation error (incommensurability of the grid-averaged model outputs with the point observations) and observation error, but the first of these is likely to dominate. Currently the predominant practice in climate science is to invoke the caveat

*Department of Mathematics, University Walk, Bristol, UK, mailto:j.c.rougier@bristol.ac.uk

"conditional on the climate model being correct", which is very unsatisfactory for both statisticians and policy-makers, and, one hopes, for climate scientists too. Therefore this paper represents a major advance in technique, and also in utility.

This statistical model will be the focus of my discussion, because everything quantitative follows from it. Section 3 examines the structure of the statistical model, and section 4 the statistical diagnostics. Section 5 looks at the issue of calibration more broadly, and section 6 is a brief summary. I'd like to start, though, with a section on 'cutting feedback'.

2 Cutting feedback

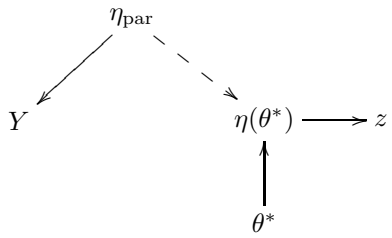
'Cutting feedback' (a term suggested by Nicky Best) involves informally restructuring the statistical model, technically a violation of coherence. In our response to the discussants in Goldstein and Rougier (2007), Michael Goldstein writes of a personal communication from de Finetti, approving of the quote

> Traditional Bayesian models can be better understood by explicitly recognizing and distinguishing between two fundamentally different meanings for probability statements. The former is the use of probability (or prevision) as the quantitative expression of the knowledge of an individual. The latter is the use of probability as a purely technical intermediary quantity helpful in translating generalized knowledge into precise statements of the former kind. (Goldstein 1981)

When making probabilistic statements about complex physical systems like climate, it is the end-product that we sign-off on: the probability that global mean temperature in 2100 is two degrees higher than today, for example. How we get there and how we document our journey, in the papers we write and the seminars we give, is an important part of establishing the authority of our assessment. But it is mistaken to think that this authority stands or falls on a simple audit of formal correctness. I'm sure we are all aware of the limitations of probability as a model for reasoning, and to insist on coherence in the development of our inference is rather like treating our climate models as perfect: something we might do as an expedient and temporary place-holder, while we develop a more nuanced approach. In this light, perhaps I over-emphasised coherence in Rougier (2007b), which outlined the basis for probabilistic reasoning about climate using evaluations of an imperfect climate model. But maybe not: unwitting incoherence is definitely to be avoided!

Cutting feedback can be illustrated in the standard 'best input' approach to model-based inference for complex systems (see, e.g., Goldstein and Rougier 2004, 2007). In the 'best input' approach, we assert the existence of a model-input θ^* such that the model evaluation $\eta(\theta^*)$ is sufficient for the system observations, z. The Directed Acyclic

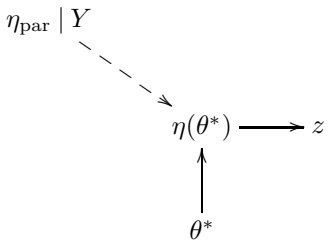
Graph (DAG) for this approach is



(1)

(neglecting a minor edge from Y to $\eta(\theta^*)$). The dashed edge is just to highlight the structure of the statistical model, particularly with reference to the Sansó *et al.* DAG below (eq. 4). Here Y is the ensemble of model evaluations, and η_{par} denotes the statistical parameters of the 'emulator', the statistical model for the physical model (O'Hagan 2006). We need an emulator for η because we cannot afford to evaluate MIT2DCM at every candidate value for θ^* (and even if we could, this might not be a sensible use of computing resources).

When we cut feedback, we collapse the edge $\eta_{\text{par}} \to Y$ into the vertex $\eta_{\text{par}} \,|\, Y$:



(2)

It is easy to show that this would follow if $Y \perp\!\!\!\perp z$, which (1) indicates is not the case (see, e.g., Pearl 2000, sec. 1.2.3). So this is technically incoherent.

In his recent Wald Lecture, Jim Berger talked about cutting feedback in terms of isolating the parts of the statistical model that we are confident about from those that we are not. In the 'best input' approach we are typically least confident about the representation of model structural error, the edge $\eta(\theta^*) \to z$. So cutting feedback in this case is about making sure that η_{par} gets its information from the high-quality data in Y, and not in questionable form from z. Of course if we judge that $\text{Var}\left(z \,|\, \eta(\theta^*)\right)$ is large, then $\Pr Y \,|\, z \approx \Pr Y$, and cutting feedback in this way seems quite natural. We cannot push this argument too far, though, because the intention is to learn about θ^* from z: this would be futile if the information from z was obscured by the model's structural error. So if we cut feedback in this way we are always going to be somewhat incoherent.

I tend to see cutting feedback more as a diagnostic tool. Once we cut feedback we can do the update $\eta_{\text{par}} \,|\, Y$ off-line, and the emulator that we construct and use in our inference can be pre-tested. The tests I find most revealing are leave-one-out (building

the emulator with all but the ith evaluation and then predicting the ith, for $i = 1, \ldots, p$) and one-step-ahead (ordering the evaluations and predicting the ith using the previous $i-1$). Rougier *et al.* (2007) provides an illustration of these diagnostics, and show how the two tests can complement each other. One-step-ahead is strongly related to Phil Dawid's Prequential approach (Dawid 1984; Cowell *et al.* 1999), and provides a simple scalar assessment of $\eta_{\text{par}} \mid Y$. This can be used for mild tuning of hyperparameters, although when committing this particular misdemeanor I prefer to use the full set of diagnostics.

This discussion of cutting feedback is not a digression, although it may seem like one. Building the emulator off-line is standard practice, which is why the theory and practice of emulation has become an important separate strand in computer experiments. Sansó *et al.* use the 'best input' approach, as shown in their eq. (1), where the unstated "assumption" (I would prefer to say "choice"), is that $\xi \perp\!\!\!\perp \{\eta, \theta^*\}$; they write $\boldsymbol{\theta}$ rather than θ^*. But they choose not to cut feedback in $\eta_{\text{par}} \to Y$, and so we are denied diagnostic information specific to the performance of their emulator (although I will suggest below that their diagnostics are in fact dominated by Y). Cutting feedback would seem to be a natural choice here, given that Y is so informative about η_{par} (426 evaluations varying only three inputs, arranged as an irregular grid). Interestingly, though, Sansó *et al.* *do* cut feedback in a different way, as I now show.

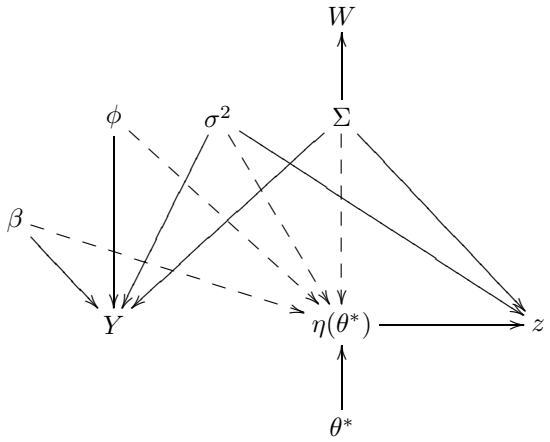
3 The ubiquity of Σ

To draw Sansó *et al.*'s DAG, we first have to clarify their emulator, to find the parameters that are identified with η_{par} in (1). The emulator is of the general form

$$\eta(x, \theta) = \sum_i \beta_i \, h_i(x, \theta) + u(x, \theta), \tag{3}$$

where h is a set of pre-specified regressors, θ is the model parameter (or input, Sansó *et al.* write t), and x the index variable of the model-output. Techniques for multivariate emulators, i.e. emulators that treat the collection $\eta(\theta) \triangleq \big(\eta(x_1, \theta), \ldots, \eta(x_n, \theta)\big)$ jointly, are still developing. Rougier (2007a) contains a short review, and presents an approach suitable when all of the outputs have the same type (so that it is reasonable for different values of the index variable x to have the same regressors and variance multipliers)—which is exactly the situation here. The statistical parameters of Sansó *et al.*'s emulator are the regression coefficients, β, plus the parameters in the variance function of u (the emulator residual) namely the correlation lengths ϕ, the variance scalar σ^2, and the variance matrix Σ. There is an additional parameter γ in Sansó *et al.*'s exposition, but

this is set to 1 in practice. The DAG is then



(4)

(again, neglecting a minor edge from Y to $\eta(\theta^*)$, which represents the contribution of the residual). Here W is an ensemble of control runs from a General Circulation Model (GCM), and dashed lines have been used to link the structure of this DAG back to that in (1). The two extra edges from Σ and the extra edge from σ^2 are innovations, to be discussed below.

This is not quite the DAG that Sansó *et al.* use, however. First, they cut feedback on Σ, by collapsing the edge $\Sigma \to W$ into the vertex $\Sigma \,|\, W$; in other words they do an off-line update of Σ using the GCM ensemble. Following the same analysis as before, this would be implied by $W \perp\!\!\!\perp \{Y, z\}$. However, this is, to my mind, rather more contentious than cutting feedback on $\eta_{\text{par}} \to Y$ in (1), because W in (4) is strongly linked to both Y and z, via Σ.

Now we turn to the interpretation of Σ, and its scalar multiplier σ^2. Σ plays three roles: it is the natural variability of the GCM, it is the variance of the emulator residual, and it is the variance of the model structural error. In line with the decision to cut feedback on $\Sigma \to W$, I am going to treat it initially as the GCM natural variability. So the questions become: (1) Is GCM natural variability a reasonable proxy (except for a scalar multiplier) for the MIT2DCM emulator residual variance? and (2) Is GCM natural variability a reasonable proxy (except for the same scalar multiplier) for the variance of the MIT2DCM structural error?

Take the MIT2DCM residual first of all. We can think of u as comprising two parts: first, the higher-order terms that were excluded from the regressors in h; second, the natural variability of the MIT2DCM model. To clarify this second point, MIT2DCM has a dynamic ocean module coupled to a dynamic atmosphere module, and the result is that the model's precise response is extremely sensitive to θ, so that there is a component of variation in $\eta(\theta)$ that has a very short correlation length. Therefore we might write

the residual as

$$u(x,\theta) = \underbrace{u_1(x,\theta)}_{\text{higher terms}} + \underbrace{u_2(x,\theta)}_{\text{nat. var.}}, \tag{5a}$$

with $u_1 \perp\!\!\!\perp u_2$. It seems reasonable to choose to treat the natural variability of MIT2DCM as some multiple σ^2 of the natural variability of the GCM, Σ. This would result in a variance function such as

$$\text{Cov}\left(u_2(x_i,\theta), u_2(x_j,\theta')\right) = \begin{cases} \sigma^2\,\Sigma_{ij} & \theta = \theta' \\ 0 & \text{otherwise,} \end{cases} \tag{5b}$$

with a nugget in θ approximating a very short correlation length. For u_1 we need to account for systematic effects in both the model parameter and the index variable. We might choose a tractable separable structure such as

$$\text{Cov}\left(u_1(x_i,\theta), u_1(x_j,\theta')\right) = r(\theta,\theta';\phi)\,\Psi_{ij} \tag{5c}$$

where r is a correlation function with parameter ϕ, and the main contribution to the variance matrix Ψ might be assessed from the lowest-order terms in x that are excluded from h, typically squared terms and first-order interactions. In contrast to (5), Sansó *et al.* have, for the whole of u,

$$\text{Cov}\left(u(x_i,\theta), u(x_j,\theta')\right) = \sigma^2 r(\theta,\theta';\phi)\,\Sigma_{ij} \tag{6}$$

some strange blend of the contributions from the two parts, not consistent with either one or the other dominating. So when we are updating our judgements about Σ it is hard to assert that we are learning about the natural variability, given that Σ is forced in this part of the statistical model to adopt some of the characteristics of Ψ, and r to adopt some of the characteristics of a nugget. If anything, we'd have to say that the data Y are dragging Σ away from being natural variability.

What about the other role for Σ, where Sansó *et al.* have

$$\text{Var}\left(z \mid \eta(\theta^*), \sigma^2, \Sigma\right) = \sigma^2\,\Sigma. \tag{7}$$

The use of scaled GCM natural variability as a proxy for the variance of the structural error is common in climate science (see, e.g., Murphy *et al.* 2004). If you are faced with a difficult elicitation, and you find you have to hand a variance matrix of the right shape, with the right units, it is undoubtedly tempting to drop it in. But GCM natural variability is a property of the GCM: it does not proxy the difference between the GCM and the climate system. In climate science this has been appreciated and discussed, but only recently has there been a genuine effort to determine a variance for the model structural error that is not based on internal variability (Murphy *et al.* 2007). In Sansó *et al.* the unnatural effect of this choice is compounded by the scaling factor σ^2 being constrained to be the same value as the scaling factor in the MIT2DCM residual variance (through imposing $\gamma = 1$). Therefore, like Y, the data z are dragging Σ away from being natural variability, but in a different direction.

The broader question of how to assess model structural error variance in practice is unresolved, particularly for models with multivariate outputs. Rougier (2007b) suggested that climate scientists use their judgement, which was dismissed by the scientists themselves as "too subjective" (praise indeed!). The method in Murphy *et al.* (2007, sec. 3g), based on the use of a multi-model ensemble (MME: a collection of different models run with the same boundary conditions and forcing), is less subjective. At Durham (UK) Leanna House is also developing an approach using a MME, structured by second-order exchangeability, as part of the MUCM project (`http://mucm.group.shef.ac.uk/`). Craig *et al.* (2001) use comparisons between low-resolution and high-resolution versions of the model, anticipating the more formal development in the reified modelling approach proposed by Michael Goldstein and myself (more on this in section 5).

To summarise, we are unsure how to interpret Σ in the Sansó *et al.* statistical model: it starts off as GCM natural variability but it is pulled one way by Y and another by z. And along with the difficulty of interpretation, there is the question of whether forcing Σ to play these three quite different roles overly constrains the statistical model. In their paper Sansó *et al.* have not defended their treatment of Σ directly, on the basis of any physical rationale, but they might still persuade with extensive diagnostic testing.

4 Diagnostics

The purpose of diagnostic testing would be to reassure us that the constraints in Sansó *et al.*'s DAG are not at odds with the data in $\{W, Y, z\}$. This is clearly different from doing a sensitivity analysis, which involves changing the marginal distributions *within* the constraints of their DAG. In their sensitivity analysis Sansó *et al.* focus on the marginal distribution of θ^*. They refer to this as 'the prior', but the prior—if it is anything at all—is the joint distribution of $\{\theta^*, \beta, \phi, \sigma^2, \Sigma\}$. The only part of this that Sansó *et al.* specify as proper is the marginal distribution of θ^*, but clearly their updated distribution could be sensitive to the marginal distribution for, say, Σ, no matter whether it is improper or proper. Each component of $\{\beta, \phi, \sigma^2, \Sigma\}$ has well-defined physical units, so diffuse but proper marginal distributions should be easy to specify, and then a simple sensitivity analysis on these choices—e.g. halving and doubling the standard deviations, once they exist—would be interesting.

Incidentally, some climate scientists have identified that one problem with a Bayesian analysis lies with 'the prior', but are under the illusion that this can be ameliorated by including evidence-based information from 'the likelihood' along with the posterior (see, e.g., Frame *et al.* 2007). Sansó *et al.*'s analysis provides a clear example of how this distinction between the properties of the prior and the likelihood is rather naïve in a complicated inference.

Sansó *et al.* present us with diagnostics based on holding-out 43 of the 426 evaluations from Y, and then predicting the model response on the hold-out and comparing it with the actual values. (As discussed in section 2, these diagnostics might have been specific to the emulator, if the emulator had been constructed off-line. As it is, they are extracted from the full update of $\{\theta^*, \beta, \phi, \sigma^2, \Sigma\}$.) However, I suspect that there is

plenty of information about MIT2DCM from the 383 evaluations that remain in Y. The experimental design for Y was a multi-level grid, so the evaluations that remain will almost certainly still do a good job of spanning the three-dimensional model-parameter space. Therefore I am not surprised that the diagnostics show that the hold-out sample is predicted well, but I am not sure that this tells us much about the statistical model for $\{\theta^*, W, Y, z\}$, or the reliability of Sansó *et al.*'s conclusions about the updated distribution for θ^*: the verdict on the statistical model from the evidence in the paper is 'unproven'.

I think holding out z and predicting it would have been more telling: in order for the information in W and Y to get to z it has to traverse the whole of the statistical model, including the various roles for Σ. For the case where two types of output are combined, it would be natural to hold-out one type and predict it with the other. Goldstein and Rougier (2006) discuss diagnostics based on predicting z, extending to the circular problem of predicting z using the model evaluated at the best guess for θ^* based on z, which is what happens, informally, when model parameters are 'tuned'.

A final observation on diagnostics. In a comment on a Royal Statistical Society read paper by John Haslett and co-authors (Haslett *et al.* 2006), I wrote

> I wonder how much we really trust the inferences that we draw from the fully probabilistic analysis. I suspect that much of the fine scale structure we observe in the results comes from trading off tail probabilities in the various model components, none of which we really believe. (Rougier 2006)

I can only imagine that this 'battle of the tails' is more pronounced when the marginal distributions are improper. Ida Scheel, a PhD student at the University of Oslo, is developing visual diagnostics to examine exactly this issue in statistical models structured as chain graphs (of which DAGs are a special case). Another solution is to switch to a Bayes linear approach (see, e.g., Craig *et al.* 1997, 2001; Goldstein and Rougier 2006), which gives me an opportunity to plug Michael Goldstein and David Wooff's new book, *Bayes Linear Statistics: Theory & Methods* (Goldstein and Wooff 2007).

5 Calibration: the bigger picture

My final point concerns the end-product of the analysis. Three parameters of MIT2DCM are treated as uncertain, to be updated using observations on climate itself, using a statistical model. I will focus on MIT2DCM's climate sensitivity, $\mathcal{S}$, and denote the 'best' value of this input as $\mathcal{S}^*$. 'True' climate sensitivity is troublesome to operationalise, but there is no doubt that it is perceived as a property of the climate system itself. It is extensively discussed in the recent IPCC report (`http://ipcc-wg1.ucar.edu/wg1/wg1-report.html`). Edwards *et al.* (2007) gives a clear summary of the concept and various attempts to estimate it. Chris Forest's previous work with MIT climate models has contributed to our current understanding of true climate sensitivity, and I am sure this paper will also have an impact.

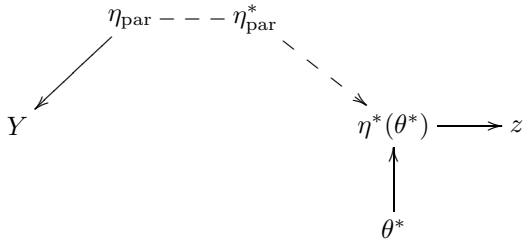
The title of Sansó *et al.*'s paper and the first line of the abstract leaves us no room for doubt:

$$\mathcal{S}^* = \text{true climate sensitivity.}$$

It is a common mistake to confuse model parameters with system values. $\mathcal{S}$ is a parameter of MIT2DCM, and as such $\mathcal{S}^*$ inherits that model's limitations. Including a careful assessment of MIT2DCM's structural error is a necessary part of updating our judgements about $\mathcal{S}^*$, but it does not turn $\mathcal{S}^*$ into true climate sensitivity: if only life were that simple! Note that I am not critical of Sansó *et al.*'s use of the 'best input' approach—this is the *de facto* standard in computer experiments, and pretty close to the state of the art—but it cannot support the interpretation they would like to give it.

Having said that, I am, in general, critical of the 'best input' approach. Michael Goldstein and I identify two main problems (Goldstein and Rougier 2007). The first one, as outlined here, is that the best input to a particular model is hard to operationalise, although we would like to define it in terms of the corresponding system property, where such a property exists. The second is that it asserts something that we may not believe, namely that the single evaluation $\eta(\theta^*)$ is sufficient for the system. A third problem, less of an issue here, is that it is demanding to combine evaluations from different models, since we would have to specify the ways in which the best inputs and structural errors of the different models were related—they cannot be independent if the models are all predicting the same thing.

The 'reified' modelling approach that we propose as a generalisation addresses these problems. We assert the existence of a 'reified model', η^*; for concreteness, think of this as a better model that we could not afford to build and evaluate. The 'best input' θ^* is an input into the reified model, and it is $\eta^*(\theta^*)$ that is sufficient for the system. Evaluations of η are useful because they tell us about the structure of η^*. The chain graph for the reified modelling approach is



(8)

(cf. eq. 1); the structure of the joint distribution for η_{par} and η^*_{par} will depend on the application. The reified approach makes θ^* more like the true system property, because η^* is a better model than η. For example, the accepted view in climate science is that the 'climate sensitivity' of a high-resolution model like a GCM is a better proxy for true climate sensitivity than the 'climate sensitivity' of an intermediate complexity model, like MIT2DCM. The reified approach also saves us from having to assert the existence of a best input for η itself. Michael and I see the introduction of the reified model as a conceptual step that clarifies the joint distribution of the actual model evaluations, and

system properties that may correspond to both model inputs and model outputs. In our rejoinder we suggest a simple way of implementing it ('direct reification', for which $\eta_{\text{par}} \rightarrow \eta^*_{\text{par}}$), in response to the suggestion from some of the discussants that reified modelling is too arduous.

If we really want to make statements about system values that correspond in some way to model inputs, then sooner or later we will have to move on from an uncritical application of the 'best input' approach. Clearly, Sansó *et al.* want to make statements about true climate sensitivity. Naturally I would be gratified if they decided to use the reified modelling approach. But they might also proceed more informally. Having updated their judgements about $\mathcal{S}^*$, they might then go on to use those judgements in some fashion, not necessarily formally probabilistic, to make some statement about true climate sensitivity. They might, at the very least, add on another chunk of uncertainty to account for the deficiencies of $\mathcal{S}^*$. To do this they would have to quantify their judgements about the limitations of MIT2DCM when used in the 'best input' approach; clearly very challenging, but if not them, who?

6 Summary

As I stated at the beginning of this discussion, I think it is splendid to have this collaboration between statisticians and climate scientists. I hope this case-study is the starting-point for a long and fruitful relationship, that will contribute methods to Statistics and results to Climate Science. I particularly commend the use of a statistical model to link model evaluations, model parameters, and system observations. This, and the inclusion of an explicit term for model structural error, are major steps forward for Climate Science.

But I do have some concerns about the implementation presented here. To summarise:

1. Sansó *et al.* choose not to cut feedback on $\eta_{\text{par}} \rightarrow Y$, denying us detailed diagnostics on their emulator performance, but they *do* cut feedback on $\Sigma \rightarrow W$, which might be considered 'excessively incoherent'.

2. In their statistical model Σ is required to play three different roles, which makes the updated distribution for Σ hard to interpret, and risks over-constraining the statistical model.

3. The emulator residual has a natural formulation that distinguishes between short and long correlation lengths in θ. The Sansó *et al.* choice, though, seems to combine these in a way that is inconsistent with with either effect dominating.

4. The choice of improper marginal distributions for $\left\{\beta, \phi, \sigma^2, \Sigma\right\}$ impedes a sensitivity analysis with respect to the marginal distributions of these parameters, and is unnecessary when all of these parameters have well-defined physical units.

5. It is always a pleasure to see statistical model diagnostics, but the ones presented

do not traverse the statistical model, and do not give us confidence that Σ can play all three of its roles. No diagnostics are presented for the joint modelling of two output types, for which additional simplifications are made.

6. The claims of the paper, notably in the title, are too bold: at best we have learnt about MIT2DCM's parameters. It is not clear, for example, what we have learnt about true climate sensitivity, or about true natural variability.

Excepting point 2, these concerns are easily addressed, from a technical point of view, although this may be time-consuming. The acid test for me is whether I would accept Sansó *et al.*'s updated distribution for θ^* as my own. At the moment I hesitate: I remain to be convinced that the three roles for Σ are compatible. For me, this would require a physically-based rationale and very extensive full-statistical-model diagnostic testing. I look forward to these eagerly.

References

R.G. Cowell, A.P. David, S.L. Lauritzen, and D.J. Spiegelhalter, 1999. *Probabilistic Networks and Expert Systems*. New York: Springer. 48

P.S. Craig, M. Goldstein, J.C. Rougier, and A.H. Seheult, 2001. Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, **96**, 717–729. 51, 52

P.S. Craig, M. Goldstein, A.H. Seheult, and J.A. Smith, 1997. Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes linear strategies for large computer experiments. In C. Gatsonis, J.S. Hodges, R.E. Kass, R. McCulloch, P. Rossi, and N.D. Singpurwalla, editors, *Case Studies in Bayesian Statistics III*, pages 37–87. New York: Springer-Verlag. With discussion. 52

A.P. Dawid, 1984. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, **147**(2), 278–290. With discussion, pp. 290–292. 48

T.L. Edwards, M. Crucifix, and S.P. Harrison, 2007. Using the past to contrain the future: How the palaeorecord can improve estimates of global warming. *Progress in Physical Geography*, **31**(5), 481–500. 52

D.J. Frame, N.E. Faull, M.M. Joshi, and M.R. Allen, 2007. Probabilistic climate forecasts and inductive problems. *Philosophical Transactions of the Royal Society, Series A*, **365**, 1971–1992. 51

M. Goldstein, 1981. Revising previsions: A geometric interpretation. *Journal of the Royal Statistical Society, Series B*, **43**, 105–130. with discussion. 46

M. Goldstein and J.C. Rougier, 2004. Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM Journal on Scientific Computing*, **26**(2), 467–487. 46

M. Goldstein and J.C. Rougier, 2006. Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association*, **101**, 1132–1143. 52

M. Goldstein and J.C. Rougier, 2007. Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*. Forthcoming as a discussion paper, currently available at http://www.maths.dur.ac.uk/stats/people/jcr/Reify.pdf. 46, 53

M. Goldstein and D.A. Wooff, 2007. *Bayes Linear Statistics: Theory & Methods*. Chichester, England: John Wiley & Sons. 52

J. Haslett, M. Whiley, S. Bhattacharya, M. Salter-Townshend, S.P. Wilson, J.R.M. Allen, B. Huntley, and F.J.G. Mitchell, 2006. Bayesian palaeoclimate reconstruction. *Journal of the Royal Statistical Society, Series A*, **169**(3), 395–438. 52

J.M. Murphy, B.B.B. Booth, M. Collins, G.R. Harris, D.M.H. Sexton, and M.J. Webb, 2007. A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philosophical Transactions of the Royal Society, Series A*, **365**, 1993–2028. 50, 51

J.M. Murphy, D.M.H. 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. 50

A. O'Hagan, 2006. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, **91**, 1290–1300. 45, 47

J. Pearl, 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press. 47

J.C. Rougier, 2006. Comment on the paper by Haslett et al. *Journal of the Royal Statistical Society, Series A*, **169**(3), 432–433. 52

J.C. Rougier, 2007a. Efficient emulators for multivariate deterministic functions. In submission, currently available at http://www.maths.bris.ac.uk/~mazjcr/OPemulator.pdf. 48

J.C. Rougier, 2007b. Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, **81**, 247–264. 46, 51

J.C. Rougier, S. Guillas, A. Maute, and A. Richmond, 2007. Emulating the Thermosphere-Ionosphere Electrodynamics General Circulation Model (TIE-GCM). In submission, currently available at http://www.maths.bris.ac.uk/~mazjcr/EmulateTIEGCM.pdf. 48

B. Sansó, C. Forest, and D. Zantedeschi, 2008. Inferring climate system properties using a computer model. *Bayesian Analysis*, **3**, 1–38. 45, 47, 48, 49, 50, 51, 52, 53, 54, 55