Low-noise projections of complex simulator output: A useful tool when checking for code errors

Jonathan Rougier[1,*], **Tamsin L. Edwards**[2], **Mat Collins**[3],
and David M.H. Sexton[4]

1. Department of Mathematics, University of Bristol
2. Department of Geography, University of Bristol
3. College of Engineering, Mathematics and Physical Sciences, University of Exeter
4. Hadley Centre, UK Met Office
* *Corresponding author,* `j.c.rougier@bristol.ac.uk`

ABSTRACT

A natural strategy to check for errors in a complex code, such as a computer simulator of weather or climate, is to make physically meaningful perturbations in the simulator parameters, and compare the results of the simulator runs against intuition. Such a strategy must balance the competing demands for CPU cycles of (i) performing many perturbations, and (ii) using long time-averages to suppress the effect of simulator noise. This paper proposes a mathematical solution which can be use to sharpen the perturbation signal in a given ensemble, namely to project the simulator output onto the column-space of linear combinations that maximise the signal-to-noise ratio. There are more refined approaches, but ours is easy to understand and to compute.

1 Introduction

This workshop concerned model uncertainty. A major source of model uncertainty is *code error*, which is not solving the equations that you thought you were solving. It is immediately apparent, from the software updates that we so frequently receive, that even the most carefully implemented codes contain errors. Large scientific coding projects such as the open-source statistical computing environment R (R Development Core Team, 2004) are fortunate in having thousands of users every day, many of whom have a clear expectation of what is correct behaviour and what is not. Consequently the core group for R has a very effective strategy for catching errors, which is to wait for them to be reported (and often fixed) by users.

Builders of specialist computer simulators for complex systems such as weather or climate, however, are not so fortunate. The codes are often so unwieldy that only a handful of users have the resources to run them. Moreover, in complex systems our intuition is often limited, and it is hard to spot an anomalous output which bears further investigation. We will see the egregious errors, of course, because the run will not complete, or its output fields will be obviously defective: all the ocean salinities might be zero, for example. But how will we spot the class of ‘middling-sized’ errors?—those large enough to make a difference to our projections, but too small to trigger our suspicions. For example, how will we spot a coding error that raises all of the temperatures over North America in 2100 by 1 $^\circ C$?

Only the most hubristic modeller would deny outright that such errors exist in a complicated code. Anyone close to the code for a large climate simulator (a General Circulation Model, or GCM) expects there to be errors—and will probably have found a few in his time—but hopes that they are all either small enough not to matter, or large enough to be egregious. Or else, with a desperate lunge at the Weak Law of Large Numbers in statistics, hopes that they all cancel out. Now that climate simulators are being used to assess quantitatively the effect of interventions such as alternative energy futures and geo-engineering, a few middling-sized code errors would render the simulator-based assessment nonsensical—except we would not know it.

If climate simulators are to be serious tools, then climate modelling groups need to have clearly-defined strategies for checking for code errors. One very obvious strategy is to perturb the 'physics' in meaningful ways, and check the code's response; in practice, this is implemented by perturbing the simulator's parameters. The fly in the ointment of this strategy, and the subject of this paper, is the presence of *simulator noise*. This is the convolution of the internal variability of the underlying equations and the truncation errors of the discrete solver. This noise can be suppressed, in situations where the forcing is time-invariant, by taking a time-average of the simulator outputs, rather than a single time-point. Therefore a modelling team with a fixed budget of CPU cycles must balance the following two competing demands:

1. More parameter perturbations to provide greater insight into simulator functioning and a greater potential to uncover coding errors;
2. Longer time-averaging to suppress simulator noise, and reveal the signal caused by the perturbed parameters.

Conventional practice in climate science is to use 20-, 30-, or 50-year averages.

This paper outlines a mathematical approach that addresses this trade-off, and allows us to extract much of the signal of the parameter perturbations without having to perform very long time-averages. In practice, many other factors will also contribute to the choice of parameter perturbations that will be run. In fact, given the expense of running a large simulator, the ensemble of runs may well be designed for an entirely different purpose, such as assessing the effect of parametric uncertainty on simulator-based projections or reconstructions. In this case it will still be very useful to have a simple approach that dampens the effect of simulator noise. But in this paper we are particularly advocating our approach for the very first stage of the analysis: comparing perturbations with intuition, to sanity-check the code.

The only inputs to our approach are a perturbed parameter ensemble (a multi-model ensemble might also do) and a long control run. A simple model of simulator signal and noise is presented in section 2, a strategy for maximising the signal-to-noise ratio is given in section 3, and for visualising the resulting simulator outputs as a map in section 4, while section 5 is a summary. The illustrations are taken from the PalæoQUMP experiment, and concern the North American mid-Holocene anomaly in the mean temperature of the warmest month (usually referred to as MTWA). These illustrations are best viewed in colour.

2 Signal and noise

We are considering a time-slice experiment, or some other type of experiment in which the forcing is changing sufficiently slowly that it is reasonable to take time-averages of the simulator outputs, to suppress simulator noise. Thus, for a specified time-interval,

$$f_i(r) = m_i(r) + \xi_i(r) \qquad i = 1, \dots, q \tag{1}$$

where $f_i(r)$ is the time-averaged simulator output i at parameter values r, $m_i(r)$ is the unknown value that would be uncovered if the run could be time-averaged over a very long time-interval, and $\xi_i(r)$ is the contribution from simulator noise, which must be treated stochastically.

A very simple statistical model for $\xi(\cdot)$ is that (i) it is independent of $m(r)$, (ii) $\mathsf{E}\{\xi(r)\} = 0$ for all r, and

$$\text{(iii) } \mathsf{Cov}\{\xi(r), \xi(r')\} = \begin{cases} \Psi & r = r' \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

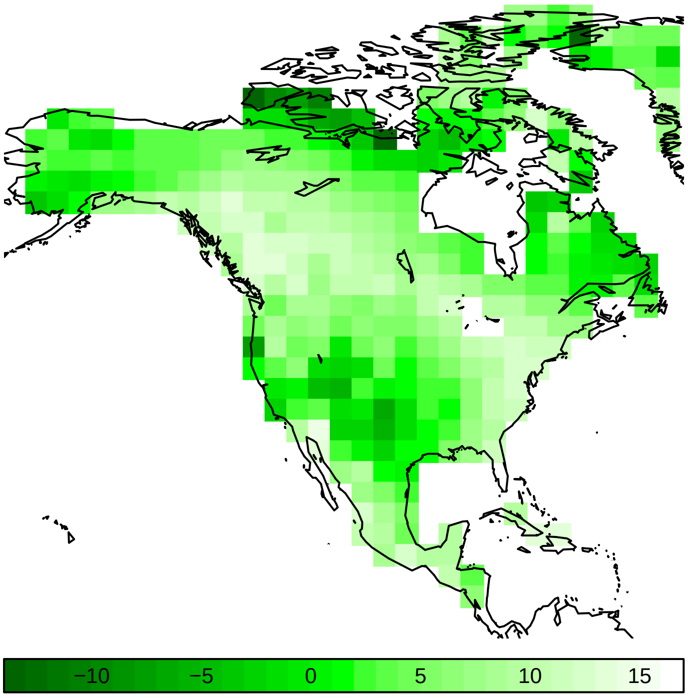

Figure 1: Signal-to-noise ratio in decibels, eq. (3)*, North American mid-Holocene anomaly, mean temperature of the warmest month. Using an ensemble of* 47 *runs from the PalæoQUMP experiment, perturbing* 32 *parameters; see Murphy* et al. *(2004) and Table 1 in Rougier* et al. *(2009), for* 31 *of the parameters, the final one is a switch between a slab and a dynamical ocean. The time-averaging was over* 20 *years.*

In other words, the mean and variance of simulator noise are invariant to the value of r, and the noise is uncorrelated across runs. This statistical model is obviously deficient, in that we fully expect internal variability, a component of $\xi(\cdot)$, to depend on the value of the parameters, r, and also boundary conditions; McWilliams (2007) refers to this as *structural instability*. But in fact we have very little evidence concerning this effect in GCMs, as the only runs we have that are long enough to estimate reliably the variance of $\xi(\cdot)$ are at the standard parameterisation of the simulator, at pre-industrial boundary conditions.

Obviously, we would like to recover $m(r)$, the signal, but we see $f(r)$, the signal plus noise. The crucial question of this paper is how to maximise the signal-to-noise ratio in our ensemble of runs. In order to consider a whole ensemble, imagine that the parameter value is a random value with some distribution over the parameter space, denoted $\tilde{r}$. Then we define the signal-to-noise ratio for output i as

$$\text{STN}_i := 10\log_{10}\frac{\mathsf{Var}\{m_i(\tilde{r})\}}{\mathsf{Var}\{\xi_i(\tilde{r})\}} = 10\log_{10}\frac{\Sigma_{ii}-\Psi_{ii}}{\Psi_{ii}} = 10\log_{10}\left(\frac{\Sigma_{ii}}{\Psi_{ii}}-1\right), \tag{3}$$

measured in decibels (dB), where the variance matrix in the numerator, $\Sigma := \mathsf{Var}\{f(\tilde{r})\}$, can be estimated directly from the perturbed parameter ensemble.

Unfortunately, if i indexes a grid-cell, this signal-to-noise ratio can be very low for a GCM with a time-average over 20 years, as shown in Figure 1. Generally, we might take anything below 10 dB as too small to be useful. Figure 1 shows a swathe running from north-west to south-east that has a useful signal-to-noise ratio, and perhaps the far south-west, but the rest of the continent has grid-cell signal-to-noise ratios of below 5 dB. Frankly, in this experiment the outputs at the grid-cell scale are dominated by simulator noise. This finding might have implications for dynamical downscaling approaches using regional climate models driven by GCM output at the boundaries.

Important note for anomalies: If the variance of simulator noise is estimated from one integration of a control run, then the estimated Ψ for an anomaly is twice this estimated variance, as there is one contribution from each integration.

3 Using linear combinations

Long time-averaging is a good way to improve the signal-to-noise ratio in simulator output, but it is often inappropriate for a perturbed parameter experiment, where a better exploration of the parameter space is the alternative. So one possibility is to use spatial averages instead, in the situation where i indexes location. We formulate this in terms of a vector of spatial weights, α, and then

$$\text{STN}_\alpha := 10\log_{10}\frac{\mathsf{Var}\{\alpha^T m(\tilde{r})\}}{\mathsf{Var}\{\alpha^T \xi(\tilde{r})\}} = 10\log_{10}\left(\frac{\alpha^T\Sigma\alpha}{\alpha^T\Psi\alpha} - 1\right). \tag{4}$$

The value STN_i is just the special case where $\alpha = e_i$, the unit vector that picks out the ith component. For any given α, STN_α is a scalar value, and it is not clear at this stage how to visualise this value; we return to this below, in section 4.

Consider the problem of choosing α to maximise the signal-to-noise ratio. This is a standard problem in multivariate statistics; see, e.g., Mardia *et al.* (1979), Appendix A.9. The answer is that if γ_1 is the first eigenvector of $Q^{-1}\Sigma Q^{-1}$ (i.e. has the largest eigenvalue), where Q is the symmetric square root of Ψ, then the maximum is found at $\alpha_1 := Q^{-1}\gamma_1$, and

$$\text{STN}_{\alpha_1} = 10\log_{10}(\lambda_1 - 1) \tag{5}$$

where λ_1 is the first eigenvalue of $Q^{-1}\Sigma Q^{-1}$. By extension of the same argument, the set of the first k transformed eigenvectors,

$$A := Q^{-1}\Gamma_1, \quad \text{where } \Gamma_1 = [\gamma_1, \ldots, \gamma_k], \tag{6}$$

say, will contain spatial averages with high signal-to-noise ratios.

One catch with this approach to maximising the signal-to-noise ratio is that it presupposes Ψ is non-singular (i.e. invertible). Indeed, we expect both Σ and Ψ to be non-singular, but our estimates, being constrained by resources, probably will not be. The simplest solution is to regularise both estimates so that they are non-singular. This type of regularisation can be implemented in many different ways, incorporating different judgements about the nature of the underlying variance matrix. We prefer to use a simple and neutral approach at this stage. A similar attitude is found in optimisation, where hessian matrices have to be constrained to be positive-definite through the stages of a numerical minimisation; see, e.g., Nocedal and Wright (2006), section 3.4. We simply add 10^{-5} to the diagonal of each estimated variance matrix, which is small relative to the scale of the individual diagonal elements. The simplest statistical justification for this approach is that the effect of the adjustment is so small that the regularised estimate of Ψ (or Σ) almost certainly still lies in the 95% confidence region for Ψ based on the estimated value.

The *scree plot* in Figure 2 shows that there is plenty of signal in our ensemble, if we know where to look: more than 60 dB in some linear combinations. Figure 3 shows the loadings of the first four linear combinations, $A_{(1)}, \ldots, A_{(4)}$. The loadings are spatially structured, which is not surprising as both the simulator output and the simulator noise are spatially coherent. However, we do not expect to be able to interpret them in terms of the simulator physics or in terms of the spatial structure of the simulator noise, since they combine these two effects. Furthermore, if there is a pattern it is likely to be mixed in the first few loadings, and then the orthogonality of Γ, which is a numerical property with no physical origin, makes the individual loadings harder to interpret.

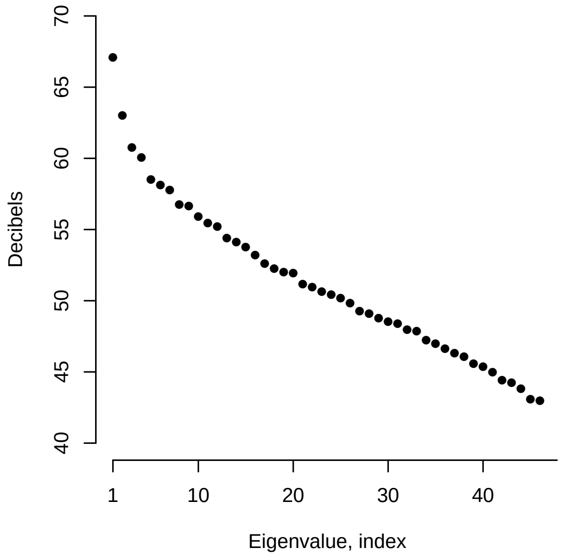

Figure 2: Scree plot of the signal-to-noise ratio of the linear combinations in A. All linear combinations after this had negligible eigenvalues.

Other approaches

Other authors in climate modelling have pursued similar approaches, although not in the context of uncovering code errors. For example, Piani *et al.* (2005) work with the same raw materials. Here we focus on Sexton *et al.* (2011), who consider the QUMP ensemble from the UK Met Office (the precursor of the PalaeoQUMP ensemble used in this paper), in the context of climate projections. In order to constrain the parameter values of their climate simulator, HADCM3, Sexton *et al.* compare the simulator output at different choices of r with measured values, to introduce a weighting scheme over the ensemble, as outlined in Rougier (2007). (Here we are simplifying slightly, by neglecting the crucial role of their *emulator*.)

For each simulator run, there are thousands of outputs to be compared with measured values, and some form of dimensional reduction is required. Sexton *et al.* use the first few variance-maximising linear combinations from the ensemble; what we would refer to as the first few eigenvectors of Σ. However, according to the analysis in this paper, their approach may well find linear combinations with moderate signal and a reasonably large component of noise; e.g. their fourth eigenvector had an emulator R^2 of only 0.65. This will lead to flatter weights across the ensemble of perturbations, and, typically, conservative projections with more uncertainty—effectively some of the simulator noise is making it through into the projections. Note that these uncertainties are not ‘wrong’; like all uncertainties they are the consequence of a particular set of choices for the statistical model. But there is an opportunity here to produce tighter projections with a small change of technique.

A more attractive option than using the eigenvectors of Σ is to use the linear combinations from a Canonical Correlation Analysis (CCA) of the simulator inputs (parameter perturbations) and the simulator outputs; for details of CCA see, e.g., Mardia *et al.* (1979, ch. 10). CCA has been used extensively in meteorology, but primarily for relating patterns in two different spatial regions, i.e. studying teleconnections (see, e.g., Nicholls, 1987). In our suggestion, linear combinations of the simulator outputs must be ‘explained’ by the parameter perturbations, and will therefore be relatively free of simulator noise. This approach does not require an estimate of the simulator noise variance. But if such an estimate is available, then it too can be incorporated through the use of an emulator. CCA provides a more principled and flexible approach to choosing the linear combinations in A than our approach, but it comes

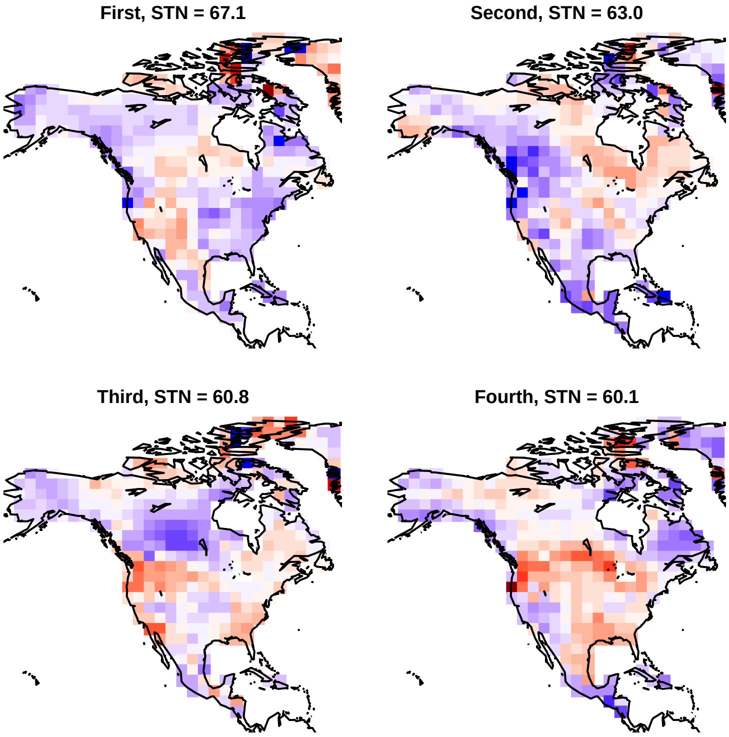

Figure 3: Loadings of the first four signal-maximising linear combinations, $A_{(1)},\dots,A_{(4)}$. These are shown with a colour-scale of blue to red through white, but they are not orthogonal or normalised, see (6).

with more statistical and computational overhead.

4 Recovering maps of simulator output

We would like to visualise the low-noise component of the simulator output, as this will approximate $m(r)$ in our decomposition (1). One way to do this is to project the simulator output $f(r)$ onto the column-space of the linear combinations with high signal-to-noise, i.e. onto the column-space of A. Formally, if A^{+} is the Moore-Penrose inverse of A, then the projection onto A's column-space is given by the matrix $P := AA^{+}$; this is discussed in textbooks on matrix theory, such as Piziak and Odell (2007, section 8.2). It is important to centre the projection, e.g. around the ensemble mean $\bar{f}$, giving

$$m(r) \approx \bar{m}(r) := \bar{f} + P\{f(r) - \bar{f}\}. \tag{7}$$

Another way to interpret the centering is in terms of the equivalent expression $\bar{m}(r) = (I-P)\bar{f} + Pf(r)$. As $I-P$ is the projection onto the orthogonal complement of the column-space of A, (7) is equivalent to filling the column-space of A with information from $f(r)$, and its orthogonal complement (where the noise is) with information from $\bar{f}$. Not to centre would imply shrinkage towards zero.

As A has full column rank, its Moore-Penrose inverse is also its left inverse, $A^{+} = (A^{T}A)^{-1}A^{T}$ (Piziak and Odell, 2007, section 4.4). Hence $P = A(A^{T}A)^{-1}A^{T}$. Eq. (7) allows us to plot any simulator run, filtering it via an operation that preserves the linear combinations with high signal-to-noise. Figure 4 shows a schematic of the procedure, with a two-dimensional simulator output and a one-dimensional projection.

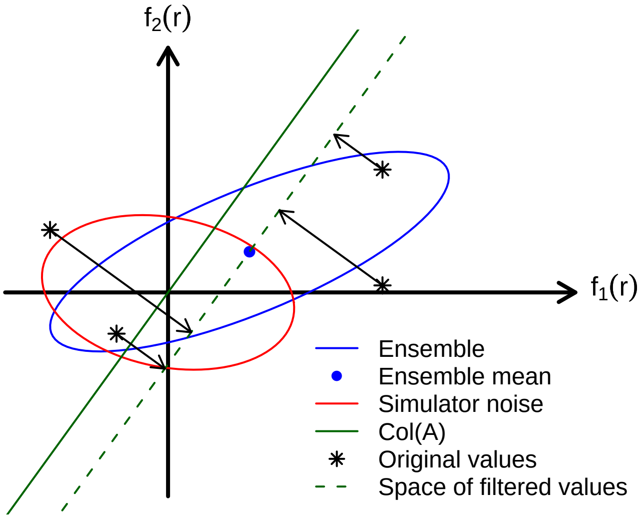

Figure 4: Schematic showing the projection onto the column-space of A in the case where the simulator output is two-dimensional, and the projection is one-dimensional (i.e. $k = 1$). The scales are chosen for easy visualisation; A in this case has a signal-to-noise ratio of only 5.5 *dB (contrast with the values in Figure 2).*

We have to decide how many linear combinations to use, k in (6). There is no reason not to keep them all if they have good signal-to-noise, as ours do, although using fewer may give similar results. Figure 5 shows the simulator run at its standard parameterisation, both the original output which includes simulator noise, and the filtered output, with three different values for k. In this case $k = 10$, the value we might have inferred from the slight dog-leg in Figure 2, seems to perform similarly to larger values. The filtered maps are clearly different from the original, with a mean reduction in temperature of about 0.5 $°C$, and changes of up to 2.5 $°C$ in some grid-cells. The heuristic explanation is that the western central large-scale positive anomaly in the original simulator output looks like a fluctuation in the simulator noise, in terms of its spatial structure, and so is filtered out (also see below, Figure 7).

Figure 6 shows simulator outputs for several different runs, ranked from smallest to largest according to the mean anomaly, with their filtered maps using $k = 40$. Unsurprisingly, the runs with the largest absolute mean anomalies tend to be filtered the most; this is an example of regression to the mean, as a plausible explanation for a large anomaly, either positive or negative, is a large fluctuation in the simulator noise.

Finally, consider again the control run (simulator parameters at their standard settings) shown in Figure 5. In this case we do actually have longer time-averages for both the pre-industrial boundary conditions (150 years) and the mid-Holocene boundary conditions (200 years). Thus we can visualise the mid-Holocene anomaly with a smaller amount of simulator noise, although the degree to which noise is reduced relative to a 20-year time-average will depend in a complicated way on the autocorrelation structure of the simulator noise, which we cannot accurately assess. Nevertheless, Figure 7 shows the result, and indicates that the filtered run does seem to have improved the representation of the simulator smooth component, at least in damping out the large positive excursion in the western central region, while preserving localised large anomalies in the north-east.

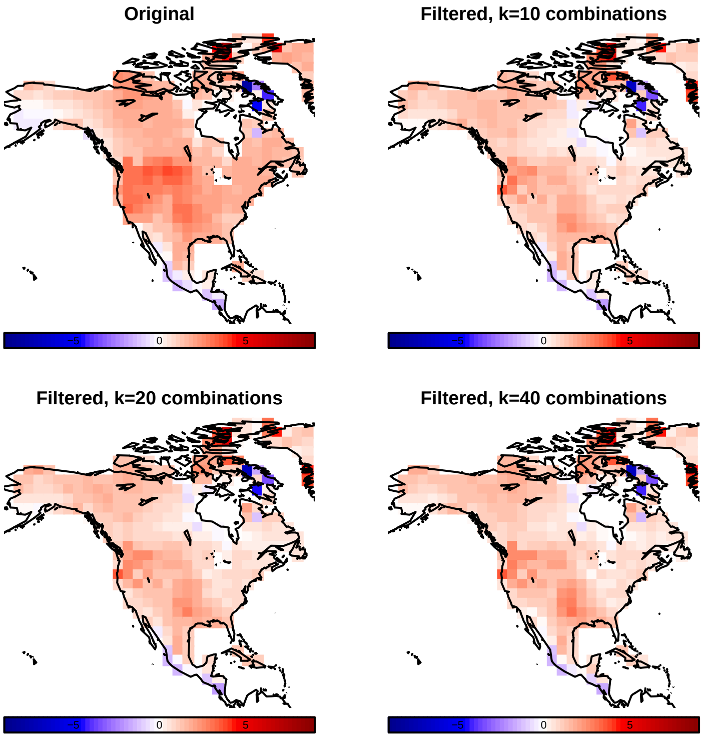
Original
Filtered, k=10 combinations
Filtered, k=20 combinations
Filtered, k=40 combinations
-5
0
5

Figure 5: Original and filtered values for the simulator run at its standard parameter settings, with three different values of k, the number of linear combinations retained. This happens to be the warmest run in the ensemble, with a mean anomaly of $+1.5\ ^\circ$*C. The mean anomaly after filtering is* $+1.1\ ^\circ$*C, for* $k = 40$.

5 Summary

Our confidence in the correct implementation of a complex code like a weather or climate simulator comes in part from checking that its behaviour conforms to our intuition. This typically involves perturbing the parameters in the simulator in physically meaningful ways, and examining the simulator response. With limited resources, however, we must trade off the number of perturbations against the length of the time-interval used to suppress simulator noise. Seldom will we have the resources to perform a sequence of perturbations where every run is time-averaged over an interval long enough to recover the smooth function underneath the noise. Furthermore, this would not be possible in experiments with time-varying forcing.

This paper has proposed a mathematical solution, using nothing but the perturbed parameter ensemble to hand, and a separate estimate of the simulator's noise variance. The idea is to keep only those linear combinations of the original simulator output that have a high signal-to-noise ratio. The filtering operation to remove the noise is to project the original simulator output onto the column-space of the linear combinations which have the highest signal-to-noise ratio.

One has to say that this approach is somewhat *ad hoc*, requiring no judgements on the part of the modeller at all, except in how many spatial averages to use in the projection, which might be assessed from the scree plot, and by comparing original and filtered simulator output. The underlying statistical model given in eq. (2) is naïve. And the approach is sensitive to the estimates of the two variance matrices: that of the ensemble and that of the control run. For a large collection of simulator outputs, possibly both estimates will be singular, and the choice here has been to regularise them in the crudest possible way. The variance of the ensemble will depend on the nature of the perturbations, and it would clearly be better to have perturbations that conform, at least approximately, to a space-filling design in the parameter region, such as a Latin Hypercube Design; see, e.g., Santner *et al.* (2003, section 5.2).

However, there is always room in a tool-box for a technique that is quick and 'not totally daft', and that is the spirit in which this approach is advanced. Establishing confidence in a complex code needs much more than this, of course, but sometimes one will get lucky—or unlucky depending on the point of view—and find a middling-sized code error with a quick and not totally daft method. This has certainly been the authors' experiences on a number of complex simulation projects.

Acknowledgements

This research was undertaken as part of the PalæoQUMP project (PI Prof. Sandy Harrison), funded by NERC under the QUEST initiative. We would like to thank members of the UK Met Office QUMP team and the PalæoQUMP team for their many contributions since the inception of the project in 2006.

References

K.V. Mardia, J.T. Kent, and J.M. Bibby, 1979. *Multivariate Analysis*. London: Harcourt Brace & Co.

J.C. McWilliams, 2007. Irreducible imprecision in atmospheric and oceanic simulations. *Proceedings of the National Academy of Sciences*, **104**(21), 8709–8713.

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.

N. Nicholls, 1987. The use of canonical correlation to study teleconnections. *Monthly Weather Review*, **115**(Feb), 393–399.

J. Nocedal and S.J. Wright, 2006. *Numerical Optimization*. New York: Springer, 2nd edition.

C. Piani, D.J. Frame, D.A. Stainforth, and M.R. Allen, 2005. Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophysical Research Letters*, **32**, L23825.

R. Piziak and P.L. Odell, 2007. *Matrix Theory: From Generalized Inverses to Jordan Form*. Boca Raton: Chapman & Hall/CRC.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3, http://www.R-project.org/.

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

J.C. Rougier, D.M.H. Sexton, J.M. Murphy, and D. Stainforth, 2009. Analysing the climate sensitivity of the HADSM3 climate model using ensembles from different but related experiments. *Journal of Climate*, **22**(13), 3540–3557.

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

D.M.H. Sexton, J. Murphy, M. Collins, and M.J. Webb, 2011. Multivariate prediction using imperfect climate models part I: Outline of methodology. *Climate Dynamics*. DOI:10.1007/s00382-011-1208-9.

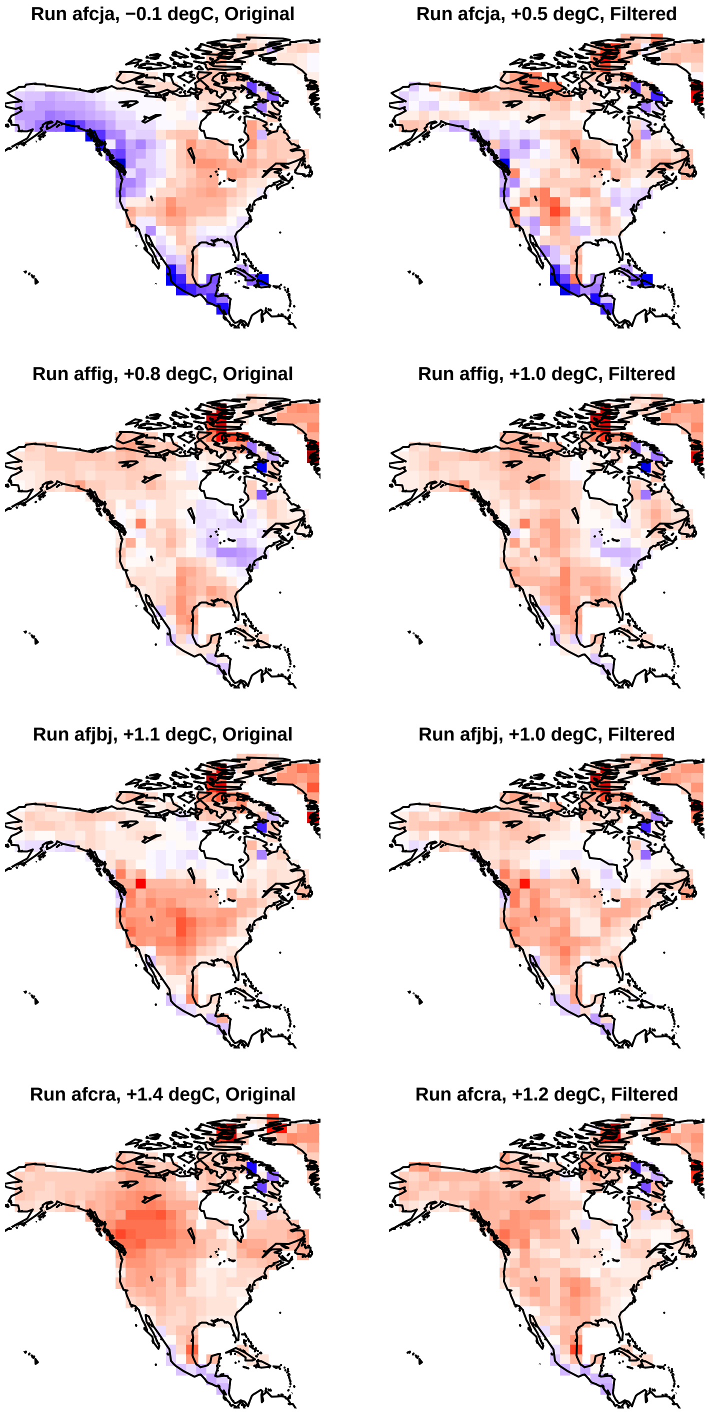
Run afcja, −0.1 degC, Original
Run afcja, +0.5 degC, Filtered
Run affig, +0.8 degC, Original
Run affig, +1.0 degC, Filtered
Run afjbj, +1.1 degC, Original
Run afjbj, +1.0 degC, Filtered
Run afcra, +1.4 degC, Original
Run afcra, +1.2 degC, Filtered

Figure 6: Original and filtered values for four simulator runs, stratified by mean temperature anomaly (shown in the panel title), with $k = 40$*. The colour scale is the same as Figure 5.*

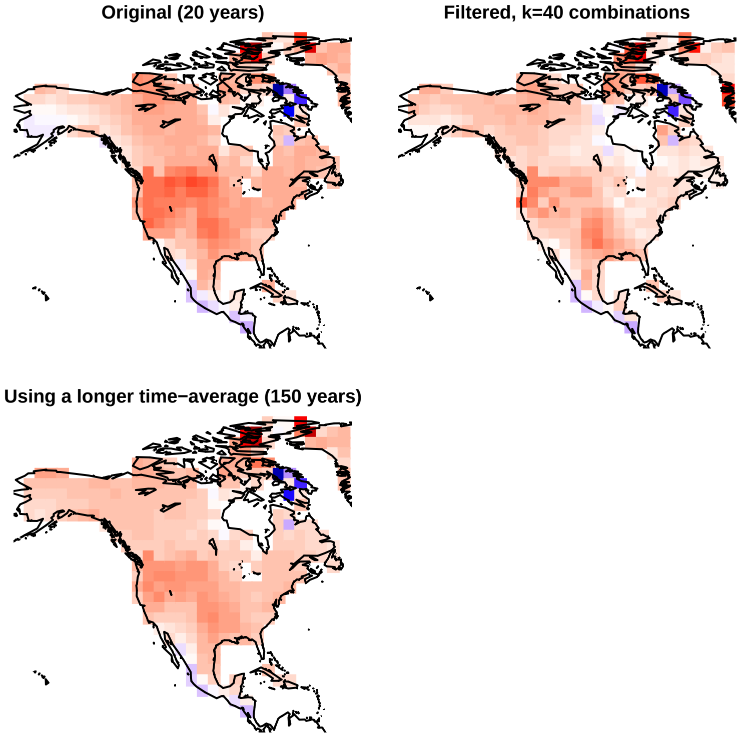
Original (20 years)
Filtered, k=40 combinations
Using a longer time−average (150 years)

Figure 7: The original and filtered maps for the simulator at its standard parameterisation, plus the original simulator output after performing a longer time-averaging. The colour scale is the same as Figure 5.