Global sensitivity analysis of the climate vegetation system to astronomical forcing: an emulator-based approach

Referee Report

September 5, 2014

1 General comments

This paper aims to take up and adapt statistical methods for global sensitivity analysis using computer simulators and apply them to an intermediate complexity climate model in order to explore the sensitivity of its response to components of the astronomical forcing on paleo timescales. A number of advances to the statistical technology are claimed, some of which have the potential to allow exploration of the sensitivity of spatio-temporal modes of variability to model parameter changes and would hence be valuable. The type of analysis advocated and any model specific conclusions that can be drawn are valuable to any field that intends to make use of that model to make inference about the behaviour of the world (either in the distant past, or going into the future). I am therefore keen that studies of this type are done and published.

However, there are a number of technical problems with the paper, as written, that must be addressed before I can recommend publication. Until the technical issues are resolved, I cannot comment on the validity of the conclusions in the paper (as they may change upon resolution). In addition to technical corrections, the paper also requires restructuring to facilitate an easier read for the intended audience. At the moment, the reader must wait for far too much already established mathematics to be re-developed in the text before they are introduced to the problem, simulator and, in particular, the important sensitivity indices that will inform them about the simulator's behaviour. This might be a valid approach for a statistics journal and a general methodology paper (though I still feel the paper would benefit from restructuring), but is not appropriate for this audience.

2 Specific major comments

1. The authors derive $S_p = \mathrm{E}_f\left[\mathrm{Var}\left[\mathrm{E}\left[f(x)|x_p\right]\right]\right]$, but are not clear in their presentation of the derivation what the target is or what the meaning of the quantity S_p actually is. To be clear, the classic sensitivity measure of interest is $V_i = \mathrm{Var}\left[\mathrm{E}\left[f(x)|x_p\right]\right]$, the expected reduction in uncertainty in the output distribution of $f(x)$ if we were to learn (or fix) x_i. I use V_i to be consistent with the wider literature on sensitivity analysis. In the paper I suggest V is reserved for these measures and emulator variances are given Σs or $c()$s as is standard. Because we cannot run $f(x)$ everywhere, we are uncertain about it. Our emulator is a characterisation of our uncertainty in $f(x)$ at unobserved points and has a posterior distribution given the ensemble. $S_p = \mathrm{E}_f\left[V_i\right]$ then is our posterior expectation of the desired quantity V_i induced by our uncertainty in model behaviour.

However, the authors want to decompose this as $S_p = S_p^m + S_p^v$ and say that S_p^v is a measure of the uncertainty introduced by using an emulator and S_p^m measures the estimated sensitivity of the simulator to the input variations. Whilst the decomposition of S_p in terms of the two pieces is fine (and follows from the derivation and the linearity of expectation), it is not clear at all what either piece actually means and I see no reason why the interpretation given by the authors should be true. If it is, it is crucial that this actually be demonstrated. The narrative of probabilistic sensitivity analysis is clean. We have sensitivity measures (the V_is and others) that we could compute if we could run the simulator everywhere. We can't, we build an emulator, we obtain a posterior expectation of V_i with respect to it. If S_p^m has the claimed property, this should be proved and it's interpretation as a posterior expectation of something, should be carefully explained (along with what it is a posterior expectation for and why it measures sensitivity). Else, all analyses and conclusions should be based on S_p only.

2. The separation of linear and non-linear effects is not helpful as I believe different things are meant by this in the climate and statistics literature. The contribution of $S_p^{m,nl}$ represents a lack of fit to a linear mean function. Whilst, in a technical, statistical sense, this represents some of the contributions to the posterior expectation of our sensitivity indices from departures from a linear fit (and we might call these non-linear effects), that language is hardly helpful in a geophysical setting, where a non-linear effect is often interpreted quite differently. To illustrate my point, conclusion 3 in section 3.4 claims that because "significant non-linearity in the North Atlantic and the Arctic" was observed, the implication is that large sea ice responses may occur and play a larger amplifying role than in other experiments without these non-linearities. The language almost implies fast changing, "snowballing" type effects. But really the non-linearities in the North Atlantic/Arctic only demonstrate that the fitted linear terms in the parameters are a poorer fit in these regions. I'll give a number of examples as to why this might happen.

   Firstly, a different linear fit may be more appropriate for the Arcitc/North Atlantic region than in the majority of the globe. The global fit in the emulator will be closer to that used required for the rest of the globe, and hence the lack of fit will be larger in the Arctic and the Gaussian process will have to do more work there. Secondly, suppose the behaviour in parameter space were piecewise linear. As one parameter increases, for the first half of it's range, the model response was perfectly linearly increasing with constant rate and, for the second half, decreasing by the same rate. Then the best linear fit would have $\beta^T h(x) \approx \alpha_0$ for some constant, and the "non-linear" term captures everything. Hence, I don't agree that the authors can claim to have identified non-linear responses in the way that they do. Or, at least, I think the language is misleading and shouldn't be here.

   Perhaps more useful would be to think of rapid and slow responses. Non-linearities are often associated with rapid change and, unfortunately, gradual change in climate studies is always termed linear (and estimated by trends). But complex models are rarely like that (slow varying quadratic or even log behaviour is common), and emulators offer the perfect chance to capture this by carefully choosing $h(x)$, so that the Gaussian process only has to capture variability over shorter correlation lengths. A partition into slowly and quickly varying effects is then possible, with the quickly varying component more akin to what is often implied by non-linearities in this literature. Whilst looking at the spatial location of regions where each part of this

decomposition of S_p will be interesting and suggestive, it would still be important to have an interpretation of these quantities in terms of a posterior expectation of something well defined. Without this, I see no good justification for claiming to have found sources of this type of variability rather than a general lack of fit.

3. The second objective the authors list in the introduction (producing geographical maps of model sensitivity/dealing with multivariate output), which they state that, to their knowledge, is not covered in the literature: is addressed in the literature and, not just in the statistics literature, but in the climate literature and for seemingly more intractable problems! The work of Ken Carslaw's group in Leeds (Carslaw et al; 2013, Lee et al. 2011, Lee et al. 2012, Lee et al 2013), some of which is published in Nature, is tackling this problem head on for cloud aerosol models and in many more dimensions of input than 3. None of this work is referenced or acknowledged and this is symptomatic of the reference to other relevant work in the rest of the paper.

   The approaches to sensitivity analysis for climate models with emulators in the literature are currently based on parallel grid box emulation to draw maps (which is an emerging trend in the field (See E. Spiller et al 2014 for similar approaches to volcano risk models)), so the authors have an opportunity to compare approaches. There are potential benefits to both, but this work is important, relevant and should be acknowledged. The current work should be compared to what exists and the benefit of taking this technically more challenging approach (if there is any) clearly laid out.

   The authors also go into great technical detail on emulator construction and multivariate approaches based on principal components. Though there is an extremely rich literature on this that is uncited, it would have been at least relevant to review what has been applied to climate models before. This may also reduce the burden of detailed mathematical development through citation. Examples include: Challenor et al 2009, Edwards et al. 2011, Sexton et al (2012), Holden et al. (2013) (for principal component emulation in climate models) and Williamson et al 2012, Williamson and Blaker 2014 (for Gaussian process emulation of multivariate output as alternatives to principal components in climate problems).

4. Experiment 20 cannot be dismissed as an outlier. The term outlier has no meaning here as the function is deterministic (even with sensitive dependence to initial conditions). As the point is "in a corner" it is outside of the convex hull of points, so it is certainly plausible that an emulator fitted on the rest of the data with that point removed would fit very badly. This is, after all, a hard test for the emulator to pass, particularly using a linear mean function. One is asking a Gaussian process which, as the simulator clearly does not have the same linear response on the edges of parameter space, is having to do all of the work to capture behaviour in the corner to extrapolate from the 26 other observations to the left out point without any guidance. The fact that it does badly and that the emulator variance increases with it's inclusion, suggests that there are parameter effects in this region of parameter space that are not captured by this emulator. The increased variance is warranted given the prior-data conflict and might be more indicative of a heavy-tailed distribution (e.g. t, see below). The solution is not to simply discard the evidence and pretend it was never seen.

   A big problem with doing this is that the emulator, without the discarded evidence

of experiment 20, still gives a prediction for the model behaviour at this setting of the parameters (and for a suitable neighbourhood around it), and we know this prediction is very very wrong, but use it anyway. Further, when the point is included, we know that the emulator will be doing better in that region of space, even though more uncertainty is included everywhere. A solution might be to run the simulator again nearby. An even better solution might be to experiment with including interactions in $h(x)$ to capture this behaviour.

The conclusion that the emulator shows that experiment 20 does not represent the simulator's response over the whole input space, which the authors draw as their final conclusion, does not follow from the evidence. In fact, experiment 20 suggests that the emulator for the model that was finally used only approximates the model well in most of the parameter space, but there is a region within which it cannot be trusted. But, for sensitivity analysis in the whole space to work, we must trust the emulator everywhere. If the model behaviour is unrealistic in this part of parameter space, this is an argument for history matching prior to conducting sensitivity analysis (Williamson et al. 2013).

5. The authors have chosen to present the design they used as an algorithm taking over 1 page of space in of itself, and with around a page of additional material used to give it context within the literature. There is no need to do this. However, if it is to be done, a great deal more must be said in justification and demonstration of the algorithm's quality than the algorithm provided "satisfactory results". What does this mean and how do you judge?

   The proposed algorithm describes a rejection sampling approach for obtaining a design that meets a constraint on a function of the model inputs that uses random uniform Latin Hypercubes to generate candidates then repeats the process 1000 times and chooses the best according to a given measure of coverage (a maximin property). Steps 1-6 of the algorithm are the same as in Vernon et al. 2010. Where step 7 could be succinctly described in a couple of sentences.

   If the authors would like to set this up as an algorithm for generating small designs in constrained parameter spaces, that is fine, but they need to go much further in justification and literature comparison than they have. They must also explore sampling properties of the resulting design. However, as this paper is not really about design, much space and reader effort could be saved by outlining the rejection sampling approach using Latin Hypercubes and referring to the existing literature, then describing this extra step taken to try to make the design more space filling than it would be otherwise.

6. The layout of the paper is not friendly to the readership and hinders understanding of the method. It reads too much like a derivation of sensitivity analysis from emulators, with far too much technical detail than it needs to; instead of an explanation of sensitivity analysis followed by the necessary machinery required in order to do it with emulators and with multivariate output. Currently the reader must wade through 5 pages on emulators and 3 pages on design before the motivation for doing any of this is clear. The paper should focus the reader on what is required in order to perform sensitivity analysis (the V_i and the V_{-i} using my notation above), explain how the statistics literature allows us to get posterior expectations for these using emulators and present the equations for any quantities needed in order to do this, and no more (e.g. equations (7) and (8) are definitely not needed here). The

authors should consider how many of these equations are required in the main text and whether some can be moved to a technical appendix. An introduction of the model and it's parameter space could come before any of this.

7. It is unclear what aspects of the proposed methodology are new for this paper. The generalisation of sensitivity analysis to principal component based emulators might be. If so, say so and please provide a little more technical justification for (20) and (21) (maybe in an appendix). This section seems throwaway, yet nothing is cited so it may be new and, if it is, it is extremely sparse compared with the detailed derivations of existing technology. At this point, a comparison of this approach with the current multi-emulator approaches described by Lee et al. (2013) is particularly important, as the authors make the simplifying assumption of only looking at the diagonal of a covariance matrix (which may lead to similar interpretation, though see below).

3 Specific minor comments

1. The theory of experimental design is not a response to being unable to fill a full factorial with enough experiments. Full factorial designs are part of experimental design!

2. Full factorials are wasteful if an input is inactive.

3. Sacks et al (1989) didn't introduce the Bayesian meta-model, nor did Kennedy and O'Hagan (2000). The former was not Bayesian and used kriging. The latter came years after the Bayesian introduction (probably due to Currin et al in 1991) and is for multi-level models so is not relevant to this point.

4. 906 line 10, reads as though cubic splines are the smoothest function GPs mimic.

5. 906 line 25, the normal approximation to the t-distribution is usually taught as accurate enough in practice for $n - q > 30$ not 10. Oakley and O'Hagan (2004) develop the theory of sensitivity analysis used here in terms of student t's. With the left out point, the number of points outside 1, 2 and 3 s.d.s might be too many for a Gaussian, but just fine for a t-process with 24 degrees of freedom. At the very least this should be discussed when going into fine detail with diagnostics.

6. The authors mention that the nugget is due to the initial condition uncertainty, yet estimate it with the penalised log likelihood even though the authors have run a (small) initial condition ensemble. Perhaps it would be worth discussing this.

7. Page 909 line 20 argues that it is shown in section 2.5.4 that there are strong computational benefits to fixing the correlation parameters for each component. I don't see that it is shown at all in that section. It is certainly addressed, but if the argument in that section does demonstrate this point, it doesn't do so clearly enough and the reader has to work far too hard. The amount of computational benefit should also be commented on.

8. A comparison of philosophies, at this point, with other spatial methods would also be relevant. Fixing the correlation parameters for every grid box in a parallel approach is common, for example. If the PC approach is viewed as more flexible (a strong case would have to be made), how does this assumption affect that argument?

9. I don't understand the need for constraint on parameter space and hence the need for a rejection sampling based design. As written, it seems that, if one wants to rule out $e > 0.05$ then $-0.05 < i_1, i_2 < 0.05$, and this can be rescaled to $[-1, 1]^3$ so that a vanilla LHC design can be used. I feel I must have missed something subtle here, but on careful reading I can't spot it and so other readers may also miss it, hence the way in which a vanilla LHC would break this constraint should be made clear.

10. Comment on only spending 27 of 81 runs on different sets of parameter values, then only using 26 of these in the sensitivity analysis. It does seem very wasteful.

11. 914-915 Why talk about the distribution of $\eta(x_p)$ at all? Yes it is a linear transformation of a Gaussian process, but the sensitivity measure of interest, $\mathrm{Var}\left[\eta(x_p)\right]$, is not. If the authors can get from $\mathrm{E}_f\left[\mathrm{Var}\left[\eta(x_p)\right]\right]$ to an equality that allows the decomposition they want later, please start from here (the current theory in sensitivity analysis, and demonstrate the novelty in the new approach).

12. Notation needs addressing throughout the paper. The best example is the use of m. This is the mean function of the emulator, a superscript, and on 919, I am genuinely unsure what is happening. We have an $m \times m$ covariance matrix and two uses of m in equation (20). All x's have been dropped on 919 so it is unclear what the notation means here. I think the m_ks and the v_ks need x's as do terms like h. Please make things as clear as possible. If notation is used once anywhere in the paper, either keep it defined as it is, or, if it is essential to re-use letters of the alphabet, make it clear that notation has changed by saying so.

13. Page 919: "The covariance matrix is thus of dimension $m \times m$". What covariance matrix? Is this the all important $\mathrm{Var}\left[\eta(x_p)\right]$? Say so.

14. Using only the diagonal of the covariance matrix (assuming above interpretation). Do the authors gain anything by using PCs rather than individual emulators for each grid box as is currently done in climate studies? Comment on the difference between the approaches.

15. The practical computation section could go in a technical appendix.

16. Page 922 point 3, Heavy tailed distributions such as student t...

17. Page 923 3.2.1 Choice of m for integration. What is m this time?

18. Figures don't appear in the order they are referred to (e.g fig 5).

19. What is the variance associated with the choice of initial conditions referred to in 3.2.2 and how is it calculated? If this is nugget variance, estimated using penalised likelihood based on an ensemble from one setting of the initial conditions (even though 2 other sets of ICs are available) this is unacceptable.

20. The principal components (u vectors) should be plotted somewhere. They will allow us to find out if spatial sensitivity conclusions have basis dependent features (suggesting that the class of spatial distributions of sensitivity we can see with this approach might be limited by the choice of basis).

4 Technical corrections

- 908 9 $f : R^d \rightarrow R^p$ with $m >> 1$!!
- p917 Line 1 missing we.