1 Response to anonymous reviewer #1

We are thankful to the reviewer for his detailed, constructive and relevant comments. We are reproducing below the original review (abridged when necessary) and inlined our responses. Quotes form the article are shaded.

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.

> **AUTHORS RESPONSE:** This comment was made by both reviewers. Our methodological and scientific objectives are better explained in the revised version. We have also reduced, where appropriate, the repetition of previously material, summarised the more technical questions and moved some of the more technical developments (in particular, the experiment design) to an appendix. The introduction was also entirely revised.

2 Specific major comments

1. The authors derive $Sp = \mathbb{E}_f \mathbb{V}\mathrm{ar}[E[f(x)|xp]]$, but are not clear in their presentation of the derivation what the target is or what the meaning of the quantity Sp actually is. To be clear, the classic sensitivity measure of interest is $Vi = \mathbb{V}\mathrm{ar}[E[f(x)|xp]]$, 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 $c()$ as is standard [...].

   > **AUTHORS RESPONSE:** We certainly agree that notions are difficult in this subject and there was considera room for improvement after the first version. We largely adhered to the suggestions of the reviewer.

2. 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 Spm measures the estimated sensitivity of the simulator to the input variations Whilst the decompo- sition of Sp 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 ....

   > **AUTHORS RESPONSE:** The reviewer is correct: this follows from Oakley and O'Hagan (2004). All the discussion now uses the previously noted S_p (now V_p in the new notation). This said, the integral Σ_{tot}, i.e., the emulator covariance averaged over the whole input space, may

still be used as an *informal* measure of the order of magnitude associated with using an emulator rather than the simulator itself.

> In particular, the quantity Σ_{tot} may be used as an informal measure of the amount of variance that is being introduced by the emulator as as surrogate to using the actual simulator: we expect this quantity to remain *small* compared to the quantities of interest T_p ad V.

3. 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. [. . .]

   AUTHORS RESPONSE: It is certainly true that the "linear" and "non-linear" dichotomy is used and abused. The reviewer #1 is right about pointing out the possible confusion between 'slow vs fast' (fast climate response response to slow astronomical changes) vs other effects that may be associated with non-linear behaviours such as, e.g., non-monotonous response. Let us think again about the objectives of the paper: demonstrate the potential of an emulator-based approach of sensitivity analysis to astronomical forcing—possibly to be used later with much more computationally demanding models—to address typical problems of palaeoclimatology. We adopt here a hypothesis of time-scale separation, i.e., assume that the ocean-atmosphere system is quasi-stationary with respect to the astronomical forcing and thus, as correctly identified by the reviewer, our interest eventually is in detecting regions of the input space that correspond to steep gradients in the climate response. Following these considerations and this reviewer's comment we chose to focus on simulating the "maximum rate of change" of climate variables, estimated by using as input realistic successions of astronomical forcing that can be computed from the available astronomical solutions.

4. 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. [. . .]

   AUTHORS RESPONSE: There were indeed references lacking. We have included the appropriate references. In this context it is probably best to emphasis the global sensitivity analysis process (Lee et al., 2011-2013) , PCA analysis (Holden et al., 2010, GRL), calibrating an Earth-Model of Intermediate Complexity (Edwards et al. 2011) or inferences about climate sensitivity (Rougier et al., 2009). We note also that Schmittner et al. 2011 have compared different solutions for emulation in a palaeoclimate context. A few words about the PCA vs independent emulators. The PCA and the independent emulator approach are two different ways of building multivariate emulators, but there are many other possibilities (coregionalised emulators for example). Certainly the PCA approach is unlikely to be going too far wrong if we are emulating enough of the components, so that the sum of the eigenvalues for these components is close to the total sum. It is likely to produce maps which are smoother and capture spatial structure better than the independent emulator approach (as PCA is designed to exactly capture this spatial variation), as the independent emulator approach will produce different sensitivity indices for each grid box, with no requirement for smooth variation in the these indices. But we don't see one method as inherently better or worse than the other, but for any given situation one approach could outperform the other. For example, the idea of using same length-scale for every emulator is questionable. We provided a sensitivity analysis comparing same lenth-scale with per-PC optimum lenth-scales for the PCA emulator, but it would be impossible to do detailed checks of 2000 independent emulators, so this assumption is more worrying for independent emulators than for the PCA emulator. Furthermore the method of using PCA emulators has been around for a while (at least since Higdon et al 2008) - the PCA emulator isn't the new part of this paper. Following these comments we added a the brief discussion and. We also now use the output covariances to provide *fingerprints* of obliquity and precession effects.

> Now that the notation and relevant concepts are introduced, the potential advantages and drawbacks of the PC emulator over the independent emulator approaches may briefly be summarised.
>
> - The PC emulation is based on the calibartion of n' Gaussian process models per output variable. As we will see, we use $n' = 10$. We explained that computational cost may be saved by using the same lengthscales for all Gaussian processes, though in practice computational cost remains affordable even when using independently optimised length scales. So, for example, the impact of the same length-scale assumption may be assessed much more easily than it would be with 2048 emulators (i.e., the number of grid points) for each output.
> - The PC emulator implicitly assumes a covariance structure of model outputs, equal to the covariance of the outputs of the experiment design. This may effectively reduce the emulator posterior variance, especially if experiment output is noisy (e.g.: short averages in a model with high interannual variability). However, this may also exaggerate the dependency of the analysis and intepretation on the experiment design.
> - Finally, the PC emulator provides co-variance indices between any two simulator ouputs. It therefore provides a picture of the spatial structure of the simulator response to individual and combined factors.

5. 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).

AUTHORS RESPONSE: Correct, we no longer use the term outlier and mention that this experiment is badly predicted.

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 cor- ner 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 its 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 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).

AUTHORS RESPONSE: History matching may not be the relevant concept here because, again, we not considering the inputs as something uncertain for which we would need to establish a posterior. Let us now come to the main point. It is now explained that experiment 20 oscillates at low frequency, at first sight similar to the dynamics of Dansgaard-Oeschger events. The oscillation period varies as the deep ocean temperature adjusts to the forcing and at the end of the experiment it is roughly 800 years. Consequently, a 500-year average (as used here) may randomly produce different fields. The emulator, founded on the hypothesis that the data supplied for calibration represent a stationary climate, has no chance to capture this. In fact, the climate system oscillates in this experiment 20 between two meta-stable states, which may be termed as “warm North Atlantic” and a “cold North Atlantic”, and the emulator predicts the warm equilibrium (attached figure). After the original submission of the manuscript we tried a couple of more experiments with nearby input configurations, and realised that the oscillator behaviour occurs for these parameter choices as well. The response to be given this state of affairs is a matter of judgement. This is where we have the possibility to some expert argument from the climate science side : does the oscillator behaviour of exp. 20 matter for the conclusions we want to draw about the response of climate to astronomical forcing? We could, for example, start an iterative experiment plan procedure, the objective of which would be to delineate the 3-D contours of the experiment design within which the oscillatory behaviour occurs, in order to build an emulator specific to that zone. This is a paper on its own, and the added value of this work for palaeoclimate science is questionable, for two reasons. The first one is that the region of the parameter space where this occurs is in fact almost never reached by Nature. In particular, the obliquity of 22° used for experiment 20 is in the lower 0.013th lower percentile (i.e., values lower that 22 ° occur less than 0.013 % of the time) according to the Laskar et al. 2004 solution over the last 10 million years. So it has in fact no weight on the global sensitivity measures. The second reason is that the oscillator behaviour has been documented before in Goosse and Renssen 2002, and Loutre et al. 2014 and generally judged to be 'non-robust' : i.e., the parameter region where this oscillation could occur is sensitive to physical parameter changes well within uncertainties. Therefore, we prefer to acknowledge the behaviour, consider that it *could be of significance* (i.e. we are definitely *not* ignoring the experiment or pretending that it never occurs) but use discard it from the emulator design procedure to the benefit of the performance of the emulator on the rest of the experiment domain, which covers quasi 100 % of the actual astronomical forcing region. The text now includes the following paragraph:

> Experiment 20 is the lowest configuration of obliquity (22°). This is lower that any actual obliquity during the Pleistocene (the minimum being 22.07° for Laskar et al. 2004). In this configuration, LOVECLIM develops a slow oscillation pattern that may be reminiscent of Dansgaard-Oeschger oscillations: millennial transitions between a warm and a cold North Atlantic phase, with fast warming and slow cooling (*Reference here to new figure; see Response to Reviewer #2).* The phenomenon, a known feature of LOVECLIM (Gossse et al. 2002; Loutre et al. 2014), can be described as the apparition of a cold North-Atlantic phase that is being visited stochastically and increasingly frequently as obliquity decreases. According to a couple of experiments not further discussed here, this cold phase is being visited shortly once during the entire experiment at obliquity of 22.5 ° (lowest 7th percentile), though obliquity threshold is likely to depend on the configuration of precession. The oscillation itself could be of physical relevance for past climate variability, but the limits of the phase space region in which the oscillation occurs are likely to be sensitive to uncertain parameter changes. At this stage, one possible response would then be to identify, by sequential experiment design, the region of occurrence of the phenomenon and develop an emulator specifically aimed at characterise this oscillation. Given the likely sensitivity of the oscillation on model parameters, the significance of this enterprise for palaeoclimate interpretation is unsure. We rather choose to ignore the experiment for the time being (the following diagnostics ignore experiment 20), but briefly discuss the possible consequences of this choice in the final discussion.

6. 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.

   > **AUTHORS RESPONSE:** The reviewer probably refers to the article of Willamson and Vernon "Efficient uniform designs for multi-wave computer experiments" found on the arXiv currently submitted to JASA. To our defense we were not aware of this article when preparing our own design. Also, there is very little literature that we are aware of on non-cubic designs (remember we need to exclude high eccentricities, so step 4. was certainly not accounted for in any previously published material) and we believe that this is a relevant point for future work on palaeoclimate emulation. This said, we agree that this experiment design and can be reduced and best placed to an appendix, and that we do not present enough evidence about the effeciency or superiority of our algorithm agains alternatives to warrant so much space in the main text.

7. The layout of the paper is not friendly to the readership and hinders understand- ing 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 expla- nation 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.

> **AUTHORS RESPONSE:** We have considered this comment of the reviewer and rewritten the paper thoroughly.

8. 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).

> **AUTHORS RESPONSE:** We cannot of course claim to have developed a *new* kind of emulator, but we believed reasonable to use an experiment design with LOVECLIM, an Earth Model of Intermediate Complexity, as a test case to study the potential applications of emulation in a palaeoclimate context. In addition, we are considering an input space that cannot be considered as "uncertain". It was explored predictably by Nature and it seems to us to be a novel application in the global sensitivity analysis literature. To some extent, we are advocating a change in current approaches of palaeoclimate modelling: besides previous attempts at "reproducing a rapid climate change at the right time" (e.g., Claussen et al. 1999), we believe that state-of-the-art methodologies coming from the science of experiment design and analysis with computer simulations may be used to scan the whole input space in order to detect where and when rapid climate change may occur in response to smooth astronomical forcing changes. The interest of this exploratory work is perhaps best shown by the reviewers' discussion: cases similar to that of the "experiment 20" (which behaved unexpectedly) or the debate between PCA-emulation vs independent emulation are likely to occur again in the future. Regarding the PC analysis, the equations (20) and (21) have to our knowledge not been published before, and we provide further details on their derivation. We also introduce "fingerprints", that is, eigenvectors of co-variance matrices, to characterise the effects of precession and obliquity.

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!

> **AUTHORS RESPONSE:** That is correct, though again let's consider the context of palaeoclimate modelling, largely based on factorial designs, our objective being to make it clear to the audience that thinking carefully about the design is a wise investment. The section on experiment design is shortened and the point is made more concisely.

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

> **AUTHORS RESPONSE:** That is correct, but we failed to understand what the reviewer suggests here.

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.

> **AUTHORS RESPONSE:** Sacks did use a meta-model strategy and considered appropriate experiments design. We do not claim that it was Bayesian. The reference to Kennedy and O'Hagan was superfluous in this context, and best compensated for by well-chosen citations about the recent applications in the climate context, suggested by the reviewer.

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

> **AUTHORS RESPONSE:** Indeed. Corrected

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.

> **AUTHORS RESPONSE:** We reproduced the barplots using the Student distribution, the visual result is hardly distinguishable from the first version. In fact, due to the nature of the PCA emulator, the variance at any grid point will be the *sum* of several student-t distributed quantities (the distribution of which has to be estimated by Monte-Carlo), plus a variance associated with residual PCs that is Gaussian distributed. The result is certainly quite close from being Gaussian, and it might well be that the original barplots were more exact, since they did not require Monte-Carlo simulations.

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.

> **AUTHORS RESPONSE:** In the original version of the manuscript, we wrote:
>
> > The nugget term, $\nu \mathbb{I}_{i=j}$, was originally introduced to account for measurement errors in geospatial data analysis (Cressi et al. 1993) In emulators, it may also be introduced and justified, either as a regularisation *ansatz* to avoid poor matrix conditioning Pepelychev, 2010, as a way to account for non-explicitly specified inputs (in the present case: initial conditions, sampling time and length), or as a way to account for the mis-specification in the correlation function (Gramacy and Lee, 2012)
>
> We are thus not claiming that it only represents initial conditions uncertainty, nor do we use it to quantify this uncertainty, and hence we are unsure of the way to change our text here.

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.

AUTHORS RESPONSE: The original manuscript was explaining this (italics added for this version):

> These vectors may be efficiently computed in the particular case where all the components of the PC emulator use the same parameters Λ and ν. Indeed, as observed after equation (18), the terms $\hat{\beta}$ and $\boldsymbol{E}$ are independent of $\boldsymbol{x}$. Hence, integrals only need be carried over $\boldsymbol{hT}'$, $\boldsymbol{TT}'$, and $\boldsymbol{hh}'$. As the latter are independent of the calibration data, they can be taken out of the summation and only need to be computed once. *We thus gain a computational factor of 50, compared to a situation where these integrals would be computed for all possible combinations of k, k'.*
> Likewise, the calibration data $\boldsymbol{y}$ enter the posterior variance only through $\hat{\sigma}^2$, which is independent of $\boldsymbol{x}$. Again, the triple integrals need be computed only once for all the components of the emulator if the correlation parameters are constant across the different PCs.

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?

   AUTHORS RESPONSE: see the above discussion about PCA vs independent emulators.

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 $0.05 < i1, i2 < 0.05$, 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.

   AUTHORS RESPONSE: The input factors are indeed first scaled, this was written but made even more explicit now. The key point regarding the rejection is that it is based on a *combination* of i_1 and i_2, i.e., the rule is $i_1^2 + i_2^2 < 1$. We clarified the description of the algorithm.

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.

    AUTHORS RESPONSE: There is a misunderstanding here. There are 27 different sets of parameter values, and each parameter value is used in three simulations: twice with active vegetation, but different initial conditions (and we could verify that the results converge to quasi the same equilibrium, except exp. 20 which does not converge to a quasi-stationary state), and once with inactive vegetation. This seemed to us the most natural way of controlling the effects of vegetation and initial conditions. In fact, the number '81' never appears in the text and we are unsure of what to add to make things more clear.

11. Notation needs addressing throughout the paper. The best example is the use of m . . .

    AUTHORS RESPONSE: References to earlier works have allowed us to reduce the amount of maths. In addition, we have reduced as much as we could the introduction of dummy variables as well as the use of the $\mathbb{E}$ and $\mathbb{V}$ar symbols, and made use of symbols T and V classically used in the global sensitivity analysis literature, and avoided the re-use of letters. We agree with the reviewer that there was considerable scope for improvement.

12. Page 919: The covariance matrix is thus of dimension $m \times m$. What covariance matrix? Is this the all important Var[?(xp)]? Say so.

> **AUTHORS RESPONSE:** As we now use the number p rather than m (overleaded) for the number of output, the covariance matrix *of the emulator for LOVECLIM.* We clarified the point
>
> > The covariance matrix of the emulator for LOVECLIM is thus of dimension $p \times p$ and provides information on the joint uncertainty of any two the simulator outputs. In practice, the computation of point-wise sensitivity indices only require to know the diagonal of this matrix.

13. 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.

> **AUTHORS RESPONSE:** The question of PC-emulator vs individual grid-box emulators is discussed above. The scientific conclusions seem insensitive to this choice, but we gain computing time, and we believe less worrying to use (and test) constant parameters in the PC emulators that with independent emulators for each grid box. Furthermore, we now make full use of the covariance matrix : we term "fingerprints" the eigenvectors of the covariance.

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

> **AUTHORS RESPONSE:** Agreed.

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

> **AUTHORS RESPONSE:** Yes, but we have clarified the fact that using the exact student distribution provided by the emulator theory does not provide a suitable fix here.

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

> **AUTHORS RESPONSE:** The original manuscript included the definition : "The sample size m is chosen empirically to be large…", so that one was pretty clear, but we agree that there was a general problem of notation overloading in the manuscript that we have addressed.

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

> **AUTHORS RESPONSE:** Fixed

18. 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.

> **AUTHORS RESPONSE:** The reviewer may be reassured. There are two sets of initial conditions available (for active vegetation), so we simply averaged, for all points of the design, the point-wise square difference between the outputs obtained from the two experiments. This was the very reason of using two sets of initial conditions. The text is clarified
>
> > Variances explicitly associated to inputs (precession and obliquity) largely dominate both the variances associated with using the PC emulator, and the variance associated with the choice of initial conditions (estimated here by comparing, for all deisgn points, the outputs obtained with the two sets of experiments).

19. 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).

AUTHORS RESPONSE: There is a dilemma here. To be comprehensive this would represent 30 plots that would barely be discussed individually. Not great for the clarity of the paper, the more so that we otherwise show that the PC emulator produces results that are very similar to those obtained with the individual emulators. Hence, show three principal components in supplementary material, the rest in digital form, provide a variance analysis for each PC, and added this brief discussioni in the text

> The variance analysis reveals a mixture of precession and obliquity effects on each principal component. The principal component analysis is thus not effective in separating the effects of astronomical forcing. The reason is simple: the fingerprints of precession and obliquity are not orthogonal, and thus cannot be readily recovered by principal component analysis of the model outputs.

A canonical correlation analysis would also be effective in separating input factors, but its use for an emulator is more intricate due, precisely, to the non-orthogonality of the components. Similar conclusions are obtained with GDD and annual precipitation.

1 Response to anonymous reviewer #1

We are thankful to the reviewer for his detailed, constructive and relevant comments. We are reproducing below the original review (abridged when necessary) and inlined our responses. Quotes form the article are shaded.

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.

> **AUTHORS RESPONSE:** This comment was made by both reviewers. Our methodological and scientific objectives are better explained in the revised version. We have also reduced, where appropriate, the repetition of previously material, summarised the more technical questions and moved some of the more technical developments (in particular, the experiment design) to an appendix. The introduction was also entirely revised.

2 Specific major comments

1. The authors derive $Sp = \mathbb{E}_f \mathbb{V}\mathrm{ar}[E[f(x)|xp]]$, but are not clear in their presentation of the derivation what the target is or what the meaning of the quantity Sp actually is. To be clear, the classic sensitivity measure of interest is $Vi = \mathbb{V}\mathrm{ar}[E[f(x)|xp]]$, 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 $c()$ as is standard [...].

   > **AUTHORS RESPONSE:** We certainly agree that notions are difficult in this subject and there was considera room for improvement after the first version. We largely adhered to the suggestions of the reviewer.

2. 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 Spm measures the estimated sensitivity of the simulator to the input variations Whilst the decompo- sition of Sp 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 ....

   > **AUTHORS RESPONSE:** The reviewer is correct: this follows from Oakley and O'Hagan (2004). All the discussion now uses the previously noted S_p (now V_p in the new notation). This said, the integral Σ_{tot}, i.e., the emulator covariance averaged over the whole input space, may

> still be used as an *informal* measure of the order of magnitude associated with using an emulator rather than the simulator itself.
>
> > In particular, the quantity Σ_{tot} may be used as an informal measure of the amount of variance that is being introduced by the emulator as as surrogate to using the actual simulator: we expect this quantity to remain *small* compared to the quantities of interest T_p ad V.

3. 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. [. . .]

> **AUTHORS RESPONSE:** It is certainly true that the "linear" and "non-linear" dichotomy is used and abused. The reviewer #1 is right about pointing out the possible confusion between 'slow vs fast' (fast climate response response to slow astronomical changes) vs other effects that may be associated with non-linear behaviours such as, e.g., non-monotonous response. Let us think again about the objectives of the paper: demonstrate the potential of an emulator-based approach of sensitivity analysis to astronomical forcing—possibly to be used later with much more computationally demanding models—to address typical problems of palaeoclimatology. We adopt here a hypothesis of time-scale separation, i.e., assume that the ocean-atmosphere system is quasi-stationary with respect to the astronomical forcing and thus, as correctly identified by the reviewer, our interest eventually is in detecting regions of the input space that correspond to steep gradients in the climate response. Following these considerations and this reviewer's comment we chose to focus on simulating the "maximum rate of change" of climate variables, estimated by using as input realistic successions of astronomical forcing that can be computed from the available astronomical solutions.

4. 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. [. . .]

> **AUTHORS RESPONSE:** There were indeed references lacking. We have included the appropriate references. In this context it is probably best to emphasis the global sensitivity analysis process (Lee et al., 2011-2013) , PCA analysis (Holden et al., 2010, GRL), calibrating an Earth-Model of Intermediate Complexity (Edwards et al. 2011) or inferences about climate sensitivity (Rougier et al., 2009). We note also that Schmittner et al. 2011 have compared different solutions for emulation in a palaeoclimate context. A few words about the PCA vs independent emulators. The PCA and the independent emulator approach are two different ways of building multivariate emulators, but there are many other possibilities (coregionalised emulators for example). Certainly the PCA approach is unlikely to be going too far wrong if we are emulating enough of the components, so that the sum of the eigenvalues for these components is close to the total sum. It is likely to produce maps which are smoother and

capture spatial structure better than the independent emulator approach (as PCA is designed to exactly capture this spatial variation), as the independent emulator approach will produce different sensitivity indices for each grid box, with no requirement for smooth variation in the these indices. But we don't see one method as inherently better or worse than the other, but for any given situation one approach could outperform the other. For example, the idea of using same length-scale for every emulator is questionable. We provided a sensitivity analysis comparing same lenth-scale with per-PC optimum lenth-scales for the PCA emulator, but it would be impossible to do detailed checks of 2000 independent emulators, so this assumption is more worrying for independent emulators than for the PCA emulator. Furthermore the method of using PCA emulators has been around for a while (at least since Higdon et al 2008) - the PCA emulator isn't the new part of this paper. Following these comments we added a the brief discussion and. We also now use the output covariances to provide *fingerprints* of obliquity and precession effects.

> Now that the notation and relevant concepts are introduced, the potential advantages and drawbacks of the PC emulator over the independent emulator approaches may briefly be summarised.
>
> - The PC emulation is based on the calibartion of n' Gaussian process models per output variable. As we will see, we use $n' = 10$. We explained that computational cost may be saved by using the same lengthscales for all Gaussian processes, though in practice computational cost remains affordable even when using independently optimised length scales. So, for example, the impact of the same length-scale assumption may be assessed much more easily than it would be with 2048 emulators (i.e., the number of grid points) for each output.
> - The PC emulator implicitly assumes a covariance structure of model outputs, equal to the covariance of the outputs of the experiment design. This may effectively reduce the emulator posterior variance, especially if experiment output is noisy (e.g.: short averages in a model with high interannual variability). However, this may also exaggerate the dependency of the analysis and intepretation on the experiment design.
> - Finally, the PC emulator provides co-variance indices between any two simulator ouputs. It therefore provides a picture of the spatial structure of the simulator response to individual and combined factors.

5. 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).

AUTHORS RESPONSE: Correct, we no longer use the term outlier and mention that this experiment is badly predicted.

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 cor- ner 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 its 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 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).

> **AUTHORS RESPONSE:** History matching may not be the relevant concept here because, again, we not considering the inputs as something uncertain for which we would need to establish a posterior. Let us now come to the main point. It is now explained that experiment 20 oscillates at low frequency, at first sight similar to the dynamics of Dansgaard-Oeschger events. The oscillation period varies as the deep ocean temperature adjusts to the forcing and at the end of the experiment it is roughly 800 years. Consequently, a 500-year average (as used here) may randomly produce different fields. The emulator, founded on the hypothesis that the data supplied for calibration represent a stationary climate, has no chance to capture this. In fact, the climate system oscillates in this experiment 20 between two meta-stable states, which may be termed as "warm North Atlantic" and a "cold North Atlantic", and the emulator predicts the warm equilibrium (attached figure). After the original submission of the manuscript we tried a couple of more experiments with nearby input configurations, and realised that the oscillator behaviour occurs for these parameter choices as well. The response to be given this state of affairs is a matter of judgement. This is where we have the possibility to some expert argument from the climate science side : does the oscillator behaviour of exp. 20 matter for the conclusions we want to draw about the response of climate to astronomical forcing? We could, for example, start an iterative experiment plan procedure, the objective of which would be to delineate the 3-D contours of the experiment design within which the oscillatory behaviour occurs, in order to build an emulator specific to that zone. This is a paper on its own, and the added value of this work for palaeoclimate science is questionable, for two reasons. The first one is that the region of the parameter space where this occurs is in fact almost never reached by Nature. In particular, the obliquity of 22° used for experiment 20 is in the lower 0.013th lower percentile (i.e., values lower that 22 ° occur less than 0.013 % of the time) according to the Laskar et al. 2004 solution over the last 10 million years. So it has in fact no weight on the global sensitivity measures. The second reason is that the oscillator behaviour has been documented before in Goosse and Renssen 2002, and Loutre et al. 2014 and generally judged to be 'non-robust' : i.e., the parameter region where this oscillation could occur is sensitive to physical parameter changes well within uncertainties. Therefore, we prefer to acknowledge the behaviour, consider that it *could be of significance* (i.e. we are definitely *not* ignoring the experiment or pretending that it never occurs) but use discard it from the emulator design procedure to the benefit of the performance of the emulator on the rest of the experiment domain, which covers quasi 100 % of the actual astronomical forcing region. The text now includes the following paragraph:

> Experiment 20 is the lowest configuration of obliquity (22°). This is lower that any actual obliquity during the Pleistocene (the minimum being 22.07° for Laskar et al. 2004). In this configuration, LOVECLIM develops a slow oscillation pattern that may be reminiscent of Dansgaard-Oeschger oscillations: millennial transitions between a warm and a cold North Atlantic phase, with fast warming and slow cooling (*Reference here to new figure; see Response to Reviewer #2).* The phenomenon, a known feature of LOVECLIM (Gossse et al. 2002; Loutre et al. 2014), can be described as the apparition of a cold North-Atlantic phase that is being visited stochastically and increasingly frequently as obliquity decreases. According to a couple of experiments not further discussed here, this cold phase is being visited shortly once during the entire experiment at obliquity of 22.5 ° (lowest 7th percentile), though obliquity threshold is likely to depend on the configuration of precession. The oscillation itself could be of physical relevance for past climate variability, but the limits of the phase space region in which the oscillation occurs are likely to be sensitive to uncertain parameter changes. At this stage, one possible response would then be to identify, by sequential experiment design, the region of occurrence of the phenomenon and develop an emulator specifically aimed at characterise this oscillation. Given the likely sensitivity of the oscillation on model parameters, the significance of this enterprise for palaeoclimate interpretation is unsure. We rather choose to ignore the experiment for the time being (the following diagnostics ignore experiment 20), but briefly discuss the possible consequences of this choice in the final discussion.

6. 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.

   > **AUTHORS RESPONSE:** The reviewer probably refers to the article of Willamson and Vernon "Efficient uniform designs for multi-wave computer experiments" found on the arXiv currently submitted to JASA. To our defense we were not aware of this article when preparing our own design. Also, there is very little literature that we are aware of on non-cubic designs (remember we need to exclude high eccentricities, so step 4. was certainly not accounted for in any previously published material) and we believe that this is a relevant point for future work on palaeoclimate emulation. This said, we agree that this experiment design and can be reduced and best placed to an appendix, and that we do not present enough evidence about the effeciency or superiority of our algorithm agains alternatives to warrant so much space in the main text.

7. The layout of the paper is not friendly to the readership and hinders understand- ing 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 expla- nation 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.

> **AUTHORS RESPONSE:** We have considered this comment of the reviewer and rewritten the paper thoroughly.

8. 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).

> **AUTHORS RESPONSE:** We cannot of course claim to have developed a *new* kind of emulator, but we believed reasonable to use an experiment design with LOVECLIM, an Earth Model of Intermediate Complexity, as a test case to study the potential applications of emulation in a palaeoclimate context. In addition, we are considering an input space that cannot be considered as "uncertain". It was explored predictably by Nature and it seems to us to be a novel application in the global sensitivity analysis literature. To some extent, we are advocating a change in current approaches of palaeoclimate modelling: besides previous attempts at "reproducing a rapid climate change at the right time" (e.g., Claussen et al. 1999), we believe that state-of-the-art methodologies coming from the science of experiment design and analysis with computer simulations may be used to scan the whole input space in order to detect where and when rapid climate change may occur in response to smooth astronomical forcing changes. The interest of this exploratory work is perhaps best shown by the reviewers' discussion: cases similar to that of the "experiment 20" (which behaved unexpectedly) or the debate between PCA-emulation vs independent emulation are likely to occur again in the future. Regarding the PC analysis, the equations (20) and (21) have to our knowledge not been published before, and we provide further details on their derivation. We also introduce "fingerprints", that is, eigenvectors of co-variance matrices, to characterise the effects of precession and obliquity.

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!

> **AUTHORS RESPONSE:** That is correct, though again let's consider the context of palaeoclimate modelling, largely based on factorial designs, our objective being to make it clear to the audience that thinking carefully about the design is a wise investment. The section on experiment design is shortened and the point is made more concisely.

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

> **AUTHORS RESPONSE:** That is correct, but we failed to understand what the reviewer suggests here.

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.

> **AUTHORS RESPONSE:** Sacks did use a meta-model strategy and considered appropriate experiments design. We do not claim that it was Bayesian. The reference to Kennedy and O'Hagan was superfluous in this context, and best compensated for by well-chosen citations about the recent applications in the climate context, suggested by the reviewer.

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

> **AUTHORS RESPONSE:** Indeed. Corrected

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.

> **AUTHORS RESPONSE:** We reproduced the barplots using the Student distribution, the visual result is hardly distinguishable from the first version. In fact, due to the nature of the PCA emulator, the variance at any grid point will be the *sum* of several student-t distributed quantities (the distribution of which has to be estimated by Monte-Carlo), plus a variance associated with residual PCs that is Gaussian distributed. The result is certainly quite close from being Gaussian, and it might well be that the original barplots were more exact, since they did not require Monte-Carlo simulations.

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.

> **AUTHORS RESPONSE:** In the original version of the manuscript, we wrote:
>
> > The nugget term, $\nu\mathbb{I}_{i=j}$, was originally introduced to account for measurement errors in geospatial data analysis (Cressi et al. 1993) In emulators, it may also be introduced and justified, either as a regularisation *ansatz* to avoid poor matrix conditioning Pepelychev, 2010, as a way to account for non-explicitly specified inputs (in the present case: initial conditions, sampling time and length), or as a way to account for the mis-specification in the correlation function (Gramacy and Lee, 2012)
>
> We are thus not claiming that it only represents initial conditions uncertainty, nor do we use it to quantify this uncertainty, and hence we are unsure of the way to change our text here.

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.

AUTHORS RESPONSE: The original manuscript was explaining this (italics added for this version):

> These vectors may be efficiently computed in the particular case where all the components of the PC emulator use the same parameters Λ and ν. Indeed, as observed after equation (18), the terms $\hat{\beta}$ and $\boldsymbol{E}$ are independent of $\boldsymbol{x}$. Hence, integrals only need be carried over $\boldsymbol{hT}'$, $\boldsymbol{TT}'$, and $\boldsymbol{hh}'$. As the latter are independent of the calibration data, they can be taken out of the summation and only need to be computed once. *We thus gain a computational factor of 50, compared to a situation where these integrals would be computed for all possible combinations of k, k'.*
> Likewise, the calibration data $\boldsymbol{y}$ enter the posterior variance only through $\hat{\sigma}^2$, which is independent of $\boldsymbol{x}$. Again, the triple integrals need be computed only once for all the components of the emulator if the correlation parameters are constant across the different PCs.

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?

   AUTHORS RESPONSE: see the above discussion about PCA vs independent emulators.

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 $0.05 < i1, i2 < 0.05$, 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.

   AUTHORS RESPONSE: The input factors are indeed first scaled, this was written but made even more explicit now. The key point regarding the rejection is that it is based on a *combination* of i_1 and i_2, i.e., the rule is $i_1^2 + i_2^2 < 1$. We clarified the description of the algorithm.

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.

    AUTHORS RESPONSE: There is a misunderstanding here. There are 27 different sets of parameter values, and each parameter value is used in three simulations: twice with active vegetation, but different initial conditions (and we could verify that the results converge to quasi the same equilibrium, except exp. 20 which does not converge to a quasi-stationary state), and once with inactive vegetation. This seemed to us the most natural way of controlling the effects of vegetation and initial conditions. In fact, the number '81' never appears in the text and we are unsure of what to add to make things more clear.

11. Notation needs addressing throughout the paper. The best example is the use of m ...

    AUTHORS RESPONSE: References to earlier works have allowed us to reduce the amount of maths. In addition, we have reduced as much as we could the introduction of dummy variables as well as the use of the $\mathbb{E}$ and $\mathbb{V}$ar symbols, and made use of symbols T and V classically used in the global sensitivity analysis literature, and avoided the re-use of letters. We agree with the reviewer that there was considerable scope for improvement.

12. Page 919: The covariance matrix is thus of dimension $m \times m$. What covariance matrix? Is this the all important Var[?(xp)]? Say so.

AUTHORS RESPONSE: As we now use the number p rather than m (overleaded) for the number of output, the covariance matrix *of the emulator for LOVECLIM*. We clarified the point

> The covariance matrix of the emulator for LOVECLIM is thus of dimension $p \times p$ and provides information on the joint uncertainty of any two the simulator outputs. In practice, the computation of point-wise sensitivity indices only require to know the diagonal of this matrix.

13. 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.

    AUTHORS RESPONSE: The question of PC-emulator vs individual grid-box emulators is discussed above. The scientific conclusions seem insensitive to this choice, but we gain computing time, and we believe less worrying to use (and test) constant parameters in the PC emulators that with independent emulators for each grid box. Furthermore, we now make full use of the covariance matrix : we term “fingerprints” the eigenvectors of the covariance.

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

    AUTHORS RESPONSE: Agreed.

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

    AUTHORS RESPONSE: Yes, but we have clarified the fact that using the exact student distribution provided by the emulator theory does not provide a suitable fix here.

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

    AUTHORS RESPONSE: The original manuscript included the definition : “The sample size m is chosen empirically to be large...”, so that one was pretty clear, but we agree that there was a general problem of notation overloading in the manuscript that we have addressed.

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

    AUTHORS RESPONSE: Fixed

18. 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.

    AUTHORS RESPONSE: The reviewer may be reassured. There are two sets of initial conditions available (for active vegetation), so we simply averaged, for all points of the design, the point-wise square difference between the outputs obtained from the two experiments. This was the very reason of using two sets of initial conditions. The text is clarified

    > Variances explicitly associated to inputs (precession and obliquity) largely dominate both the variances associated with using the PC emulator, and the variance associated with the choice of initial conditions (estimated here by comparing, for all deisgn points, the outputs obtained with the two sets of experiments).

19. 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).

AUTHORS RESPONSE: There is a dilemma here. To be comprehensive this would represent 30 plots that would barely be discussed individually. Not great for the clarity of the paper, the more so that we otherwise show that the PC emulator produces results that are very similar to those obtained with the individual emulators. Hence, show three principal components in supplementary material, the rest in digital form, provide a variance analysis for each PC, and added this brief discussioni in the text

> The variance analysis reveals a mixture of precession and obliquity effects on each principal component. The principal component analysis is thus not effective in separating the effects of astronomical forcing. The reason is simple: the fingerprints of precession and obliquity are not orthogonal, and thus cannot be readily recovered by principal component analysis of the model outputs.

A canonical correlation analysis would also be effective in separating input factors, but its use for an emulator is more intricate due, precisely, to the non-orthogonality of the components. Similar conclusions are obtained with GDD and annual precipitation.