Review of Bounceur et al. (ESDD 4 901-943)

In this work, the authors introduce an emulation-based global sensitivity analysis to astronomical forcing of a climate model. After a review of the methodology on emulators for univariate and multivariate output and the measures of global variance, they design a set of physically consistent experiments. Finally, they perform the analysis on several variables of climatological interest to understand which inputs mostly affect the outputs, and compare this uncertainty with the variability of the emulator.

This work could be publishable, but it needs to be vastly restructured to focus more on the science rather than the methodology, and it must address some concerns about the scientific conclusions. Further, the mathematical notation needs to be greatly revised and I urge the authors to check more carefully its consistency across the work before the next submission.

General comments

- The manuscript discusses an application of the global sensitivity analysis to astronomical forcing in a climate model, yet the vast majority of the paper reviews material in emulator and global sensitivity measures literature. This is also clear from the Introduction, where only the last paragraph describe the scientific aim of the work. Since this paper does not present novelty in the methodology (or so it seems from how it is presented), the discussion of the results need to be expanded and it has to be more clear how the scientific findings add to the current science.

- I don't find the characterization of experiments 20 and 27 as "outliers" quite compelling. In the manuscript, there seems to be in two overlapping justifications:

  - different equilibrium climate for different initial conditions,
  - poor emulator performances.

  I don't think either of the points can qualify these two experiments as outliers. If there is convergence to different equilibriums, then this must be further investigated. As the authors state in the supplement:

  > It is unclear whether these patterns reveal distinct attractors of the ocean circulation states, reached from the different initial condition sets, or whether they correspond to weakly connected regions of the attractors that have randomly been sampled from the 500-year sampling and averaging procedure use for output processing.

As for the second point, a poor performance of an emulator only means your statistical modeling is overly simplistic. Therefore this indicates that, as it is, the emulator does not work well for a comprehensive global sensitivity analysis which, by definition, has to cover all parameter space.

- I found section 2.5 quite hard to read, and the mathematical notation not very clear. When introducing the global variance measures, there are three sources of uncertainty:

  - $\mathbf{x}_{\bar{p}}$, whose density is $\rho(\mathbf{x}_{\bar{p}} \mid \mathbf{x}_p)$
  - $\mathbf{x}_p$, whose density is $\rho(\mathbf{x}_p)$
  - The Gaussian process model.

  A better notation would put the pedix p or $\bar{p}$ in expected values and covariances to make it explicit with respect to which of the three sources of uncertainty you are integrating.

Specific comments

- p.903 l.20. A variance in the input factors is a consequence of assuming that the inputs are random variables, which might or might not be a reasonable assumption. In any case, it is worth pointing out your choice, and why you made this choice.
- p.904 l.12. An emulator is a computationally cheap **stochastic** approximation to the simulator. A deterministic approximation can just be a linearization of some primitive equations in the climate model.
- p.904 l.14-18. I am not sure I agree with the term "feasible here". The smooth character of the response or the correlation in the outputs can be incorporated in some emulators, and this certainly is an appealing property, but it has to do with the emulator flexibility, but not its feasibility.
- p.904 l.28-29. If you want to produce geographical maps, you have to deal with spatially correlated output, which can be regarded as multivariate.
- p.905 all equations are missing the punctuation.
- p.905 l.10-11. There is nothing Bayesian in assuming f has a probabilistic distribution. A Bayesian model elicit priors for $\boldsymbol{\beta}$ and σ^2, while a frequentist model assumed them to be fixed and unknown.
- p.905 l.24-26, p.906 l.1-2. The use of the term "global" here is problematic. $\tilde{m}(\cdot)$ is a mean response across all possible uncertainties assumed by the statistical model, and $\tilde{V}(\cdot, \cdot)$ encodes the deviation from this mean behavior. Why did you call it "global"? Further, the stochastic component is does not have to be "smooth": the random field can be non-differentiable.

- p.906 l.7. $\mathbf{y}$ represents the data and it has a likelihood, not a prior. Only the statistical parameters involved in the analysis have a prior, assuming you are working in a Bayesian framework.
- p.906 l.8. The first matrix A should be in bold. Also why are you referring to it as a Gram matrix?
- p.906 l.11. Prior distribution of what?
- p.906 l.17. I believe the correct expression is

  $$m(\mathbf{x}) = \mathbf{h}(\mathbf{x})'\hat{\boldsymbol{\beta}} + \mathbf{T}(\mathbf{x})'\mathbf{A}^{-1}(\mathbf{y} - \mathbf{H}\hat{\boldsymbol{\beta}}),$$

  so there was a missing hat in the $\boldsymbol{\beta}$ and $\mathbf{T}(\mathbf{x}^*)$ should be $\mathbf{T}(\mathbf{x})$.
- p.907 l.3-4. See the point above about the use of the term "global".
- p.907 l.11. Please use a more appropriate reference. The Matérn model is used in so many different areas that a reference to a standard textbook is perhaps more appropriate. See e.g. Stein (1999) or more recently the Handbook of Spatial Statistics. Also, the citation needs to be put in parenthesis.
- p.907 l.18. The use of nugget is necessary in the case of a squared exponential correlation function, as otherwise the statistical model would assume an overly smooth change of the output with respect to the input. Under the squared exponential, a knowledge of the output for an arbitrarily small interval in the input implies that the output is uniquely determined everywhere. So in this context, this can be seen as a form of mis-specification.
- p.907 eq (7). I believe the authors took the notation from equation (3) in Andrianakis and Challenor (2012). That, however, was a marginalized likelihood. What you wrote is a marginalized loglikelihood so please fix the notation accordingly (there is no proportionality on the log scale).
- p.908 l.3. Why did you choose $\varepsilon = 1$? Also, since this is the same symbol used to define the obliquity at p.910, l.2, it would be probably best to write equation (8) without $\bar{M}$ and ε, which are not used in the rest of the paper anyway.
- p.908 l.9. I believe there is a typo: $f : \mathbb{R}^d \to \mathbb{R}^m$.
- p.909 l.6. I would urge some caution in the use of the term "information" in this context. SVD finds the direction of maximum variance, and it could be a desirable property, but since we have a spatially index field, I am not convinced that the spatial information (i.e. the spatial correlation) of the data is minimized. SVD does not account for the spatial nature in this context and it has been shown to be potentially discarding substantial information.

- p.910 l.19. If i_1 and i_2 both depend on the eccentricity and perihelion, how can you claim they are independent?
- p.912 l.3. Please use a less colloquial beginning of phrase.
- p.912 point 3. Dividing $[-1, 1]$ in N equal width intervals and using permutations implies that you are assuming i_1, i_2 and i_3 are all equally important for LOVECLIM, at least a priori, and that every part of the parameter space is equally important. Is that a reasonable assumption?
- p.914 l.9. What is $\mathcal{X}_p$? It was never defined.
- p.914 l.18. There is a comma missing.
- p.914 l.22-23. If you assume a nugget effect in you emulator, even an infinite number of model runs would still give you an uncertain estimate of $f(\mathbf{x})$ for every $\mathbf{x}$.
- p.915 l.24. $\eta(\mathbf{x}_p)$ is a linear functional.
- p.915 l.2. The notation looks wrong. Either you use V_p or V_{pp} everywhere.
- p.915 l.6. I believe (12) should be

  $$V_{pp}(\mathbf{x}_p, \mathbf{x}_p^*) = \int_{\mathcal{X}_{\bar{p}} \times \mathcal{X}_{\bar{p}}} V(\mathbf{x}, \mathbf{x}^*) \rho(\mathbf{x}_{\bar{p}} \mid \mathbf{x}_p) \rho(\mathbf{x}_{\bar{p}}^* \mid \mathbf{x}_p^*) \, \mathrm{d}\mathbf{x}_{\bar{p}} \, \mathrm{d}\mathbf{x}_{\bar{p}}^*.$$

  Besides, this is not a variance, as stated in line 11. This is a function of two arguments, so I don't understand why it is denoted as $\mathbb{V}\mathrm{ar}_f(\eta(\mathbf{x}_p))$.
- p.915 l.7-8. I don't understand why you claim that the expectation and variance are computed with respect to the Gaussian process model for f. (11) and (12) are integrating with respect to $\mathbf{x}_{\bar{p}}$, which reflects your uncertainty on your input parameters space, and not on the emulator.
- p.915 l.16. How did you define $\mathbb{V}\mathrm{ar}(\eta(\mathbf{x}_p))$? You only defined $\mathbb{V}\mathrm{ar}_f(\eta(\mathbf{x}_p))$. I believe that here you are computing the expectation of (12) with respect to the emulator uncertainty.
- p.916 l.2. What is $\rho(\mathbf{x}_p)$? Before you only defined $\rho(\mathbf{x})$ and $\rho(\mathbf{x}_{\bar{p}} \mid \mathbf{x}_p)$. Intuitively I'd say it's

  $$\rho(\mathbf{x}_p) := \int_{\mathcal{X}_{\bar{p}}} \rho(\mathbf{x}) \, \mathrm{d}\mathbf{x}_{\bar{p}},$$

  but it was never stated.
- p.919 l.1. Shouldn't the covariance matrix be $n' \times n'$? Further, there are too many ms: one for the mean and one as a index for the sensitivity index.

- p.919 l.1-4. What do you mean by “insightful enough for our purpose”? Understanding the interaction between multiple outputs in the global sensitivity analysis seems well within the scope of this work.

- p.919 l.5. What is $\mathbf{S}_{\bar{p}}$? Did you mean $\bar{\mathbf{S}}_p$?

- p.920 l.1-14. Please rewrite this part using a more careful notation. What is $\mathbf{x} \mid \mathbf{x}_p$? Also, the inner integral is in the space of $\mathbf{x}_{\bar{p}}$, but where is $\mathbf{x}_{\bar{p}}$? Also, $\mathcal{X}_p$ is 1-dimensional, while $\mathcal{X}_{\bar{p}}$ is 2-dimensional, so I guess the integral in (22) should be 5-dimensional. Besides, what is g here?

- p.921 l.14. I guess the authors refer to “calibration” as “estimation”. In the context of computer model experiment, calibration is a completely different problem, that has nothing to do with this statistical model.

- p.921 l.14. “hyperparameters” usually describe the parameters of a prior on your parameters. You have not given any prior on $\mathbf{\Lambda}$ or ν here.

- p.921 l.27-30. Could you elaborate on the claim that lower-order principal components represent variability modes of importance? Is there any reference on this?

- p.922 l.12. “satisfactory”

- p.922 point 3. How can you claim the results are “satisfactory overall” if you noticed, correctly, that the behavior is typical heavy-tailed and you used a Gaussian emulator? To me, Figure 7 underscores a fundamental misfit of all quantities.

- p.923 l.5. “Limit” is repeated twice.

- p.923 l.6. In the previous Section you claim that your final choice of PC components is 10, yet here you use only 2. Even if parameter estimation is less demanding than Monte Carlo integrals, I am worried that your estimated sensitivity might be underestimated, as you are removing too many modes of variability from the analysis. Is the estimated sensitivity robust with respect to the increasing number of PC considered?

- p.924 l.12. A parenthesis was opened, but never closed.

- p.929 l.2-4. If the emulator cannot reproduce experiment 20 as the simulator’s response, to me it’s a sign that the emulator is wrong, not that the experiment (or the simulator) is.

- Figure ordering. Please change the figure ordering and make it consistent with the text ordering (e.g. Figure 2 is introduced in p.923, while Figure 6 is introduced in p.921).

- Figure 2. The y-axis is not informative here. I would put percentages, also to be consistent with the text.
- Figure 4, caption. "resolved" on the third line
- Figure 5. The contribution from precession is $\bar{S}^m_{1,2}$, from obliquity is $\bar{S}^m_3$ and the synergy is $\bar{S}^m_{1,2,3} - \bar{S}^m_{1,2} - \bar{S}^m_3$ as stated in p.917.
- Figures 8-9. If possible, increase the size.
- Figure 8. "across". Also, the choice of the colorbar is quite strange. It is consistent for each row, and the the bounds can differ from column to column, but choice of colors should be much more standard (i.e. blue for low values, bright red for high values).
- Supplement, p.2, 3rd paragraph. "reveals"