Review of: “On the spatial calibration of imperfect climate models” - S. Peatier, B. Sanderson, L. Terray

Summary

My comments and suggestions are based on the following interpretation of the content:

Calibration of Earth System Models (ESMs) is subject to different sources of uncertainty. Principally among these are parametric (from constant values set within the ESM) and non-parametric (the residual resulting from the ESM set to values achieving minimal parametric error). To explore these errors, the authors analyze a previously-constructed perturbed physics ensemble (PPE) of the atmosphere model in CNRM-CM6 climate model. The PPE is a set of 102 simulations from a parametric sweep across 30 model parameters. The goal of the analysis is to use this PPE to produce a small number of (new) representative parameter values whose simulations achieve a low parametric error, and display diverse behavior. The methodology is as follows (i) separate parametric and non-parametric error components displayed by the PPE (ii) dimension reduction (iii) fit an emulator for parametric-error and sample it across the parameter space (iv) threshold out values with large emulated parametric-error compared to a reference (v) find values from this subset with a large “feature” distance between them. This is tested on two example data, RMSE to surface air temperature, and MSE to five descriptive fields. The approach validation is largely through presentation of the spatial fields of the final selection of five diverse cases with similar parametric and nonparametric error decomposition .

Conclusion

I believe this is a strongly presented and interesting paper. The literature is largely comprehensive, the methods are walked through in precise detail, and the presentation of results are given in a collection of well-chosen figures. Most of the conclusions are supported by the results. I believe the scientific perspective is interesting, and a strong proof-of-concept framework is presented.

It is not a perfect manuscript however, it suffers occasionally by informal terminology, including even in the title. The conclusions also did not convince me that the results are really consequential beyond mild scientific interest. Most critically, I believe the methods within the framework are not evaluated properly; and this leads to caveats in results where the methods used for emulation are not exactly fit for purpose, and the methods to “maximize” diversity, may not necessarily do so.

I will recommend the manuscript for publication under addressing the comments below.

Major comments

- **Method deficiencies (i) the emulator:** The choice and use of emulator are not discussed, but in several places in the results it is indicated that there is room for improvement here.
  E.g. in Figure 3 - often the emulated values (that determine also the concept of "diversity" above) did not lead to close or hugely diverse simulations. I came to this conclusion as (1) the dots and triangles are often far from each other, and (2) the dots are typically more spread than the triangles, (3) one dot lead to a simulation crash for its corresponding triangle.
  Likewise in Figure 7: 4 diverse candidates were selected with emulated $p < 1$, yet all in practice most were > 1. Though this was noted in the document it was not attributed to the emulation stage being unable to well-represent the underlying model.
  Particularly in the results, the shortcomings due to the choice of emulator should be attributed clearly, and additional discussion added to the conclusions in terms of how to address this in future investigation.

- **Method deficiencies (ii) exploring diversity:** Perhaps the weakest part of the investigation involved the exploration of "diversity of solutions". Though I agree that the approach removes solutions that have similar parametric components, it is mentioned L284 that the authors "maximize" the diversity in solutions, or L177 "a selection … as diverse as possible", yet their approach is not a seemingly well defined optimization procedure. In fact the approach the authors use appears to be non-standard for such exploration, likewise with the termination of "5" final candidates is completely unmotivated. Finally the sensitivity to the choice of the threshold was also not explored. Typically for sourcing representative simulations a practitioner might use a clustering algorithms (k-means/medians/mediods etc.) over the data, and draw representative samples from the cluster centers to illustrate diversity. Such approaches are scalable and robust, and one can quantitatively assess sensitivity to e.g. the number of clusters the data appears, thresholding etc. The random removal methodology presented here appears untested. For example, if the authors ran the method over 100 trials, would they expect to find the same nodes each time, or if not, would they expect the selections of 5 display similar diversity, how are would they quantitatively measure this? How would they compare the performance of terminating at different numbers of solutions.
  Some references and/or computational validation experiments for the approach taken are needed, otherwise, again some clear admission in the methods and final discussion that this is a proof of concept, and suggest other approaches that readers can try in future to explore diversity.
  From a naive inspection of the samples in Figure 3, it does not seem out of the question that there are more representative samples one could find. I also think the use of emphatic words such as "large" diversity (L452 and elsewhere), "the key" trade-offs (L507 and elsewhere) should really be removed when one is not able to really quantify how representative these differences are, or how many trade-offs are observed due to

the selection procedure.

- **Optimality and near-optimality:** As a reader, this is the walkthrough of optimality in this manuscript.
  - An optimal solution is defined to be an exact quantity f(theta*) eqn(9).
  - This is then redefined in L158 where optimal is an input that minimized e_j.
  - Then in L168 the optimal solution is considered to be the input that minimizes an emulated objective (maybe this should be called e_{em,j}).
  - In L172 The introduction of a black-box "reference calibration" CNRM-CM (6?) is introduced as a threshold for optimal candidates (this is the first time this is mentioned in the document and it is unreferenced), it is used to rule out all candidates with greater error. It is not clear if this is calculated with real or emulated coefficients. Is the intention to have a "typical" level of accuracy, arising from current "tuning procedures"?
  - In L520 there is mention of near-optimal configurations
  - In L528 optimality and an optimal space are then unified with near optimality?

  I request that the authors solidify these notions and be consistent throughout. I also request that the authors use concrete notation for where they are using emulated vs actual scores. Finally, the threshold should be presented in a more clear fashion and explanation as to why this choice should define optimality.

- **Why explore the diversity of solutions?:** Given that the investigation is centered around exploring the diversity of these solutions of low parametric error, It was mentioned only in the final sentence as to why one wishes to do this beyond scientific curiosity. I would argue the authors should more strongly present their case throughout the text, particularly how this specific approach of selecting a few diverse candidate simulations can fit into the practitioners workflow, or how it relates to uncertainty quantification.

Minor comments

- Please change the title. Having read the document I still do not know what "Spatial Calibration" is. It is in fact this term is not once mentioned in this entire manuscript. One suggestion: "Exploration of diverse solutions from imperfect climate model calibration"?
- My own background implies that "model error" = "structural error" = "non-parametric error" = "model discrepancy" here some of these are treated differently, please be clear in defining all of these terms and ensure consistency through the text.
- L19 The literature review missed the growing works of the CliMA group: Here, for example, Idealized GCM's have been calibrated with Bayesian Formalism, using tools from data assimilation and accelerated samplers (https://doi.org/10.1029/2020MS002454, https://doi.org/10.1029/2021MS002735), the calibration approach has shown scalable to higher dimensional parameter spaces in different settings.

- L25 Stating that hand-tuning “has proven remarkably robust”, is easily misunderstood by the reader, the authors add a long caveat. Instead why not state something clear, such as “Such approaches remain popular in operational settings”
- L40 For history matching / NROY approaches also cite the studies of (https://doi.org/10.1029/2020MS002217, https://doi.org/10.1029/2020MS002225)?
- L72 This investigation explicitly explores the role of parametric error (whereas structural error often refers to an error incurred due to mis-specification of model structure, i.e. non-parametric error),
- L84 Are there findings from (Peatier et al 2022) about the validity of this PPE also be provided - it seems that results are critically linked to the exploration of this ensemble. My initial concern is that 100 members across 30 dimensions leads to poor exploration without very tight bounds and well chosen points. Perhaps the size of the non-parametric error obtained in this investigation can also shed some light here?
- L86 what is an element, why are they weighted, what is their corresponding area..? Was this sentence misplaced?
- L151 - in other applications it is also seen as a Gaussian Process (following https://doi.org/10.1111/1467-9868.00294)
- L360 “quantitatively”, should be “qualitatively”. No rigorous scoring was assigned to the amount of trade-offs, and although I am sure the “multiple optima” behavior is true, the authors did actually evaluate any optimal solutions in this investigation, rather solutions achieving below a user-chosen parametric error threshold.
- L519, L553 state that by-hand calibrations are “tractable”. This is not true, it may be the case that hand-tuned models can still be performant, but this is not the same as tractability - which additionally implies the procedure of hand-tuning to be easy, and straightforward practice, while the reality is that it is more a “dark art” of the climate modeling community.

Typos etc.

- L211 dead link
- L241-2 repeated phrase “high-order modes”
- L245 - “poitn”
- single quotes backwards e.g. L7, L522.,
- Typically latex for “theta star” has a superscript \theta^*.
- Final sentence of L16, was this meant to be here? Seems out of place