Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2018-198-RC2, 2019

Geoscientific Model Development Discussions

Interactive comment on "Improving climate model accuracy by exploring parameter space with an O(10^5) member ensemble and emulator" *by* Sihan Li et al.

Anonymous Referee #2

Received and published: 14 January 2019

Review of "Improving climate model accuracy by exploring parameter space with an O(10ˆ5) member ensemble and emulator" by Li et al.

General summary: The manuscript by Li et al applies an iterative, parameter-space refining procedure using perturbed parameter ensembles and statistical emulators to reduce regional biases of temperature and precipitation over the northwestern United States. The paper is well written and thorough, and provides a useful demonstration for reducing biases in the Hadley Centre climate models. I am support publication after the authors address the items below (listed in no particular order).

Item 1: The title seems to suggest that an ensemble of 100,000 forward simulations

was used to construct the emulators, which isn't the case. From Table 2, only a few thousand forward simulations were used. Once the emulators are trained, they can be evaluated very quickly, potentially millions to billions of times. The number of emulator evaluations O(10ˆ5) is therefore somewhat arbitrary and not significant. I recommend that the authors remove the reference to the emulator evaluations in the title because it is misleading. The authors should also consider adding information about the bias reduction goal of the study in the title (e.g. "Reducing climate model biases by exploring . . .").

Item 2: The first paragraph of the introduction describes the bias reduction goals of the study. On lines 53-54, I recommend that the authors change "simulations cast doubt on the reliability ..." to "simulations reduce the reliability ...". Later in the paragraph the authors describe prior work that found a relationship between the warm bias and shortwave radiation. This relationship motivates the need for a perturbed parameter approach (i.e. to quantify and reduce the bias). I recommend that the authors introduce PPEs in the first paragraph, rather than wait until the 5th paragraph.

Item 3: The introduction emphasizes parameter tuning as a major goal of PPEs, but using PPEs to estimate model PDFs and uncertainty is also an important application. I suggest that the authors include a statement about estimating PDFs versus parameter refinement.

Item 4: I recommend the following modification on line 146. Change "varied systematically." to "varied systematically or randomly."

Item 5: In the discussion about different types of PPE studies in the introduction, it would be worth pointing out that categories 2 and 3 may not be different from each other if a sufficient number of forward simulations are used to produce an adequate emulator over the full parameter space. With a good enough emulator, it is possible to both rule out parameter space and optimize parameter values. In this case, categories 2 and 3 are simply post-processing steps.

Item 6: Bayesian climate model calibration and MCMC are not mentioned or referenced in the introduction (e.g. Jackson et al., J. Clim. 2008), nor is optimization over multiple objectives (e.g. Neelin et al., PNAS, 2010). It would be worth including these references.

Item 7: The authors mention that little work has been done using PPEs for parameter refinement to improve RCMs. I agree with this statement, but think that it would be worth referencing prior work using PPEs for parameter refinement to improve regional climate in GCMs.

Item 8: Toward the end of the introduction, the authors describe how instead of searching for a single optimized parameter set, they consider multiple parameter sets because of the challenges of compensating errors and other effects. Rather than a limited number of parameter sets, it would be better if posterior parameter PDFs were estimated in a Bayesian sense. Doing so would understandably be outside the scope of the manuscript, though the authors should comment about the potential benefits of using parameter PDFs in their analysis.

Item 9: I would like the authors to comment about how they expect their results would differ if they swapped the order of phases 1 and 2 (i.e. first reduce biases in NWUS and then rule out regions that don't preserve energy balance).

Item 10: The authors use the so-called standard physics set (SP) as a reference point for gauging improvements from the PPE. Are the parameter values in the SP the same between the global and regional versions of the model? If so, there would appear to be a mismatch because the parameters represent unresolved processes and the standard values should be adjusted to account for differences in scale between the HadAM3P and HadRM3P. Following similar reasoning, it is difficult to see how parameter perturbations applied to the global model can be directly applied to the regional model without adjusting for scale differences. The authors should comment on and describe the implications of this potential mismatch.

Item 11: Observational uncertainty is assessed by using multiple observational datasets. For the regional bias analysis, how large are the datasets differences relative to the upscaling variability in regridding PRISM to HadRM3P?

Item 12: Presumably emulators are used for the sensitivity analysis, though this is not clear from the discussion in section 2.5. I recommend including a short summary of the emulators before the sensitivity analysis, rather than keeping all of the emulator discussion in the appendix. If emulators were used for the sensitivity analysis and are efficient to evaluate, so why not use a quantitative Sobol analysis instead of the qualitative FAST method?

Item 13: When I first read section 3.1 and interpreted the results in figure 1, I thought that the phase 2 emulator errors for SW were significant. Only later, after seeing the quality of the emulators in appendix B, did I realize that the errors were smaller than I thought. I recommend that the authors summarize the quality of the emulators earlier in the manuscript. Referring to figure 1, can the authors also comment about why there does not appear to be a very strong relationship between the LW and SW points (correlation looks like 0) and why there are no simulations in the blue ellipse with high LW?

Item 14: The OAAT relationships in figure 2 are useful as qualitative indicators of the dependence of the outputs on the inputs. While it reasonable to hold other inputs at their mean values, it would probably be more informative to use the default SP values. Can the authors regenerate the plots using SP values instead? Moreover, instead of conditioning on the values of the other inputs, it may be more useful to compute partial dependence plots that integrate over the other inputs. Unlike the OAAT plots, partial dependence plots account for interactions.

Item 15: Global sensitivity indices are presented in figure 3 using the FAST method. The text mentions that total and main effects are both computed, and that the maximum sensitivity value is 1. For nonlinear models, the sum of the total effects can be greater

than 1 because interactions are counted more than once. Can the authors clarify their statement? Also, I am wondering about the robustness of some of the differences between the sensitivity indices (e.g. for JJA-Pr). If emulators were used, can the authors estimate and include emulator uncertainty in figure 3?

Item 16: Figure 4 is highly useful, but contains a lot of information. It might be easier to digest in three separate figures (input-input, input-output, and output-output). I am also wondering about some of the differences between the phase 3 and SP values. For the input-input plots, the SP red dots tend to lie near or within the phase 3 set of points (there are exceptions for CW_land, ENTCOEF and V_CRIT_ALPHA). In the MAC-T-input and JJA-T-input plots, however, the SP and phase 3 points are completely separated from each other. It looks like the separation can be explained by ENTCOEF and V_CRIT_ALPHA, but the magnitude of the separation seems too large. For MAC-T it looks like the average difference is about 2 degrees C. But the temperature change in figure 1 is about 1 degree C when ENTCOEF is varied between 3 and 5. Can the authors explain why small changes in ENTCOEF and V_CRIT_ALPHA lead to such large (and discrete) changes in temperature? Understandably, one of the goals of the study is to reduce the temperature bias, so large changes might be expected. But the authors use an iterative and refinement strategy that I would expect to reduce the bias in small continuous steps, not the large, discrete changes that are shown.

Item 17: Given that ENTCOEF and V_CRIT_ALPHA are dominant parameters affecting temperature, it might be useful to use the emulators to further analyze and display the temperature surfaces as a function of the inputs.

Item 18: On the copy of the manuscript that I reviewed, there appears to be an artifact (i.e. a white line) at longitude 0 on all of the spatial maps (e.g. see figure 6k over Eastern Africa).

Item 19: After refining the parameter space between phases 2 and 3, the parameters that were dominant in figure 3 may no longer be dominant. Can the authors recompute

the sensitivity indices for the refined parameter space after the bias reduction?

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-198, 2018.