A learning-based method for efficient large-scale sensitivity analysis and tuning of single column atmosphere model (SCAM)
Guo et al.

In this study, the authors explore the use of a neural network-based surrogate model to find optimal parameter values for SCAM simulations of five IOPs. The conclusion is that the surrogate model out-performs SCAM when run with these optimal parameter values. Again, the paper should undergo major revision before being accepted for publication.

Major comments

In general, one of the big arguments that this paper makes is that the surrogate model is an important tool in model development because it enables a more efficient search for optimal parameter values. The authors have expanded that notion slightly by adding a discussion of optimized values computed over all 5 cases vs. individually, but have not explored potential changes in accuracy in that scenario, and have not confirmed that the improvements present in the surrogate model are matched in SCAM simulations. Beyond those topics, there is no indication that these optimal values will in fact improve global CAM simulations; if that could be shown, the impact of the paper would *greatly* increase. In general, some assertions made in the text are not supported by what's shown in the study or cited, and should be more strongly backed up. Additionally, references should again be updated and checked, and figure/table descriptions should fully describe the elements present.

Specific comments

- Line 17: Bacmeister et al. (2014) is not the correct citation for CESM, please refer to recent documentation of the full model rather than CAM alone. Similarly, the reference for CAM in the next sentence should be updated; the current citation (Dennis et al. 2012) refers to the development of a spectral element dynamical core; more recent and broader citations are available.
- Line 22: "SCAM…[is] a cheaper and more efficient alternative model for the purpose of tuning physics parameters." This discussion of SCAM and its use deserves more nuance. The authors here seem to suggest that parameter tuning can be done purely in SCAM and that SCAM is a suitable surrogate for running the full global model. While SCAM is a useful tool for initial testing of parameterizations (and perhaps tuning), it is not in fact a full substitute for running CESM globally.
- Line 29: "…has the most significant effect on results. and the Sobol method that uses the…" Sentence fragment re: Sobol method, perhaps left over from previous edits?
- Lines 67-69: These toolkits are not well introduced; if they can implement SA and tuning, what exactly is the downfall of using them? Please elaborate on what niche this is filling. The authors state that the use of more parameters and cases increase the exploration space exponentially and make the task "almost an impossible job," but don't show clear evidence of that.
    - Line 76 states that the benefit of the approach is the addition of "new SA methods in recent years." – which methods? Is that addition the only benefit (in

which case it doesn’t seem that the expansion to more cases/parameters is nearly impossible in those packages, as stated)?

- The reference to Pathak et al. (2022) is missing a DOI and is incomplete.
- Line 106: In general, the selection of PRECT is perhaps reasonable given its importance societally and the focus on convective parameterizations. Yet it seems like Pathak et al. assessed the response of a number of climate fields to their perturbations, which more closely mimics the process of model tuning. Is it not worthwhile to expand to more indicators of model performance? Near surface T/Q or surface fluxes, for instance, seem to be available for most IOPs.
- Line 110: The bounds for parameters are selected to be 50% and 150% of the values; are there not more firm limits based in the literature for what a realistic range may be?
- Lines 112-114: please indicate which fields you’ve now made tunable parameters (vs. which ones already were).
- Lines 116-118: As before, I’m still not entirely clear on what the authors mean by saying the programs needed to be recompiled and that the recompilation has enabled a larger number of concurrent instances on the Sunway supercomputer. I *believe* that this is referring to the fact that all parameters in table 3 are now namelist options, thus the model does not need to be re-built each time and can instead be cloned from a single, pre-built base case whose namelist can be modified. That does cut down on run time, but doesn’t seem related to the programs running on Sunway needing recompilation due to the supercomputer platform. Suggest clarifying that this is simply the shift of hard-coded to namelist defined variables.
- Line 174: “we decide another S=768, where NMorris is 64 and NSaltelli is 32.” What’s the sensitivity to those choices? Does 64 and 32 show some convergence to a solution?
- Line 187: “Compared to other networks, ResNet has the advantage of using less pooling.” What does this mean? Should elaborate on what pooling is first if this is an important point to make. If the following explanation is intended to elaborate on why ResNet is preferred, perhaps rephrase for clarity.
- Line 190: “ResNet18…” similar question to above – what’s the sensitivity to number of layers? Is 18 enough (and how have the authors determined that)?
- Lines 208-209: “As the sensitivity values calculated by different methods are of different orders of magnitude…” – was this expected, and arises as a by-product of using different SA methods? I.e., are the units not *all* in the form of some response driven by a change in the parameter?
- Lines 232-233: “This way, we can determine the number of parameters in the combination, taking into account the tuning effect and the amount of computation.” – Not clear what’s meant here. What is meant by ‘number of parameters in the combination,’ isn’t that set? How is the ‘tuning effect’ taken into account, as this seems like a tool that aids in tuning rather than the tuning have an effect on the process.
- Line 234: “lead to the most significance in output” – should this read the most significant *improvement* in output, or the largest increase in accuracy? Could be interpreted as the biggest change in output as written currently.

- Line 292: "…the distribution of pz4(tau)." –if zmconv_tau in figure, suggest consistent naming conventions.
    - In general, it would be more useful to use the zmconv_tau naming convention as in Fig 3 for Fig. 5 as well, to make it easier for the reader to quickly scan through.
- Figs 3 and 4 – using consistent colors in Figure 4 to represent the cases shown in Fig. 3 would increase the readability.
- Fig 4 – since the main point is the increase in *accuracy* of PRECT rather than just the change, perhaps there's some way to indicate the change in precipitation bias rather than the overall percent change? It's possible this belongs more in Figure 3 (i.e., getting a sense of how much the bias has improved given each parameter, rather than just which results are "better" than the control)
- Line 300: "Different from other cases, this case is a ocean case and is located in the Atlantic Ocean" – This seems to suggest that GATEIII is the only ocean case; does TOGAII not also target ocean? If this isn't the only ocean case, what about it being in the Atlantic makes it particularly unique?
- Table 5 –doesn't seem to add much to the paper; could this not just be said in the text?
- Lines 309-310: "It can be seen that ResNet has the best performance…" – The RMSE is indeed the lowest, and not just by a little but by quite a margin! Was this fully expected a priori? Is this consistent with any other literature?
- Figure 5: How are the sensitivity results normalized?
- Lines 322-323: "differences between the individual parameters are not as wide as in the other cases" – not sure what's meant here; that the range in parameter values is smaller? The ranges should be consistent across cases, yes?
- Line 324: "This is also consistent with its position" – by position, is the intent to say that the geographic location is more comparable? Please consider rephrasing for clarity.
- Section 4.3: The discussion of sensitivity to SA method should be expanded and clarified. It is stated that the SA results are shown, but which column is this? The Morris test seems significantly more sensitive to parameter variations than the other tests; are the others better suited for one precipitation regime vs. another? Or how is the different sequence of samples coming into play here?
- Lines 357-358: "However, the computation amount increases exponentially during the test." But isn't the benefit of using a surrogate model that you can tune more parameters at a time than using the traditional approach? If all tests for 3 parameters can be conducted in less than a minute, why not tune 4?
- Figure 7: It's unclear what the max/min stars represent or add to the plot; the min stars also aren't visible from what. I'm unsure why the y-axis is positive only; do none of the parameter combinations lead to a *decrease* in precipitation?
- Table 7: Presumably, in each case these parameters are being assigned unique values (i.e., pz4 = ?? for ARM95 vs. TOGAII); it would be helpful to include those values here as well for comparison (or reference later table).
- Line 366: "meaning average error" – should this now be RMSE?
- Table 8 is somewhat hard to interpret. Is "error" here referring to the difference in SCAM and the surrogate model? Or is it related to either's agreement with observations? Is

SCAM doing a better job simulating PRECT at all locations vs. the surrogate model (given the lower RMSE values) – that doesn't seem to be reflected in the text.

- Figure 10: Please clarify if the "model simulation output" is from SCAM or the surrogate model. It also makes more sense to plot the x-axis in time coordinates rather than timestep number.
- Lines 377-378: "The tuning of the SCAM parameters is quite productive on the time scale." – not clear what this means with the current wording. What time scale is being assessed? Does 'productive' mean increased accuracy?
- Line 381: "Although still below the observed level at about 1300 steps of TOGAII, improvement is also reflected." – I'm not sure why this timestep/case is singled out here, it seems like there are plenty of other examples of this. Similarly, the point is made that in TWP06 there are instances where control is less than observed; again, that's present in all cases, yes?
- Section 4.5: I'm not clear why the baseline and optimized models should have any different computational cost, unless baseline refers to SCAM and optimized to the surrogate model. If that's the case, that should be stated much more clearly and often. More generally, when considering overall model performance, the same model structure should be used. So of course use the surrogate model to find the optimum parameters, but to assess differences in simulation accuracy, plug those values back into SCAM. If not comparing direct SCAM cases, the accuracy gains aren't very understandable.
- Lines 388-389: "The use of NN trained surrogate models for parameter tuning can further save computational resource overhead and, in terms of results, can meet or exceed traditional optimization methods in most cases." – this may be overstated. In terms of saving computational resources, this is not illustrated by Figure 12. Figure 12 includes time in the job queue for the SCAM case; adding a bar to each case for the computational time *excluding* queue time would be needed to support this statement.
- Lines 391-392: "The model can get an enhancement in performance from 6.4%-24.4% in precipitation output," – again, I think the important part is to plug this back in to SCAM using the parameters determined by the surrogate model. Unless the suggestion is to use the surrogate model for all simulations (rather than for finding optimal parameters and model tuning), this is a critical step.
- Table 9: A better description is needed; "multi" cases undefined here.
- Section 4.6: An important point is lacking from the discussion, which is how well does the model perform (RMSE per site) when using the values found to optimize performance across all 5 sites (multi(d))? The potential benefit of regionally-varying parameters is noted, but it's not illustrated how important is it to use potentially regionally refined values vs. ones that are optimized for global performance.
- Related to above, and though this might be outside what the authors' computational resources allow – the most convincing argument for the study would be to apply the tuned parameters from the multi(d) case to a global run. That would illustrate true benefit from using a surrogate SCAM model to find optimal parameter values; without that, the potential impact of the paper is severely more limited.