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

In this study, Guo et al. explore the utility of training a neural network-based surrogate model for SCAM (single-column Community Atmosphere Model). The authors conclude that the NN model significantly improves computational efficiency without a significant loss in model performance and is thus particularly useful for parameter tuning, which is explored by studying PRECT biases in five different IOPs. While the NN model is indeed a useful tool for parameter tuning, the study should undergo major revisions prior to publication.

Major Comments

The paper relies on the premise that SCAM is time-consuming to run for these IOPs, stating that a single SCAM run can take more than hour to complete. For most user, SCAM should run *much* faster than this and take no more than 5-10 minutes at the most to complete these IOP cases. A NN surrogate model would still be significantly faster, but the benefit of it relative to hour-long SCAM cases is likely significantly overstated here. It would be nice to see more clear motivation/problems outlined in the introduction.

- Lines 217-218: "Even for the SCAM model, which takes more than one hour to finish a run, such a combined cost becomes impractical, for combined studies of multiple parameters…." - How many nodes are being used for your SCAM baseline case? Properly tuned, SCAM should run in a matter of minutes and should certainly not take more than an hour to run any of these IOP cases.
- Line 231: "With the compute time reduced to less than five percent of the original model" – is this based on a SCAM case taking >1hr to run for an IOP? I would question that baseline performance.
- Line 72-73: "Improved balance between cost and accuracy" – The accuracy of SCAM is not in question, so I don't see that the balance between cost and accuracy has significantly changed. The cost itself is also not a prohibiting factor in running SCAM, as it's already pretty efficient. The issue with assessing a larger number of parameters often lies in the tractability of analysis, not computational cost.

More generally, the paper should undergo some restructuring and refining. There are frequent references in the introduction to points that are not made until the methods (i.e., the number of samples, error improvements in variables that are not specified, etc.). Ideally, there should be a more distinct separation between the introduction (problem statement and relevant background) and the methods (specific details of the approach used and the experiments conducted here).

- Lines 164-168: "For example, the Morris sampling (Morris, 1991)… " – This statement feels like it's nearly identical to the introduction (lines 36-42). Overall, section 3.1 seems to rehash the background given in the introduction rather than describing the specific SCAM cases run.
- Line 183: "At the training stage, we reuse the 768 sets of different parameters and their corresponding total precipitation output …" - What are the 768 different sets of

parameters? How does this arise from the 11 parameters tested? Is it 768 sets per IOP? More detail is needed on the SCAM runs in order to provide context for this section.

- Section 4.1: I'm not sure how relevant this is to the results shown below; ideally, regardless of where SCAM is run, the results should be identical. Would suggest removing this full section on use of Sunway TaihuLight, particularly as it's mentioned already previously.
- Section 4.6: It's unclear to me how this differs from section 4.5; please elaborate further in the main text. Did section 4.5 not use the same range of values for all cases?
- Lines 410-411: "This is also confirmed by the experiments in the next subsection." should there be a section 4.7 here that's being referenced?
- Lines 408-409: "In addition, another scenario was considered: where the parameter configuration is the same among cases": This paragraph seemed to suggest that another experiment was conducted in which the parameters are all set to the same value in order to optimize performance across all five cases; I don't see results from that experiment though.
- Sections 4.5-4.6: I don't see an explicit discussion of cases where the optimal parameter values are carried out in SCAM rather than the NN surrogate model to confirm the results. The authors should clarify which experiments are conducted in SCAM vs. the surrogate, and consider a more explicit discussion of differences that arise when the optimal tuned parameters are used in full, online SCAM runs.

Please also ensure that figures are clearly described both in the text and the captions, and have units consistently on the y-axis, if relevant. Some of the plots also have curious signals that could be better elaborated on in the text.

- Figure 5: Could you define 'maximum fluctuation' more precisely? Is this the difference between lowest and highest PRECT value, and is the value of PRECT output hourly/daily/etc?
- Line 224: "However, the complexity of the calculation increases exponentially during the test, as shown in Table 4." I'm not sure Table 4 is the right reference here; it shows the SA methods used, but does not seem to indicate complexity or cost of these tests.
- Figure 9: Which SA method is being used in this figure? Units on the y-axis would also be helpful.
- Figure 9: TOGAII seems like PRECT is less sensitive to tau than the ARM97 case, which is at odds with the heat map in Figure 8. Is there a reason for the apparent discrepancy?
- Figure 12: Please add units to the y-axis
- Figure 13: A more detailed description of this figure in the text would be helpful. It's unclear what improvement is being plotted, and what the 'original' case in blue is given that the 'enhancement of effects' on the y-axis is due to the addition of NN and/or grid searching. Is the blue not the control case then?
- Figure 14: Similarly, more detail would be useful. What units are the overhead in computing given in? Is it obvious that the overhead should be the same for the Original and NN cases?

- Lines 421-422: “we can see that the parameter values taken between the two cases of land convection are positively correlated in the same parameter space” it’s surprising to me that the correlation is equal to 1 between ARM95 and ARM97 despite apparent differences between them in Fig 15. Could the authors elaborate on why this occurs?

Specific Comments

- Lines 1-2: “The Single Column Atmospheric Model (SCAM) is an essential tool for analyzing and improving the physics schemes of CAM.” Please specify that CAM in this case is the Community Atmosphere Model
- Lines 6-8: “By reusing the 3,840 instances with the variation of 11 parameters…” Suggest avoiding using specific numbers like this in the abstract; without an explanation, it is unclear what “the 3,840 instances” are. Either add clarification/context, or remove the specific number of instances.
- Line 15: Should this read “the *effects* of global climate change”?
- Line 18: This citation of CAM is outdated, please point to the scientific articles describing CAM instead. The title of this particular citation in the references points to CAM3 and the link itself is broken.
- Line 18: “Of these components, the Community Atmosphere Model (CAM) (UCAR., 2020), is the one with the most complexity.” It’s hard to say that more model *complexity* is contained in one model component than another; could the authors clarify/justify what’s intended by this statement?
- Line 20: “Participated in continuous numerical integration,” - Consider rephrasing for clarity; do the authors mean that in coupled climate simulations, this is a source of uncertainty?
- Line 22: “However, as a general circulation model (GCM), CAM takes a long time and a large amount of resource to run…” Given that ESM is already defined above, the authors should continue to use that notation rather than also defining GCM (unless a distinction is intended, which could be elaborated on).
- Line 24: “good alternative model” – rephrase for clarity. What’s meant by ‘good’ here – cheaper, more efficient, etc?
- Line 25: What is meant by SCAM only needs “one process”?
- Lines 28-29: “Sensitivity analysis (SA) is a method for investigating how uncertainty in the model output is assigned to the different sources of uncertainty in the model input factors, and the participants (Saltelli et al., 2010).” It’s not clear what the participants are; please clarify.
- Line 41: “quasi-random sequence by Sobol (Sobol’, 1967) and other researchers),” please specify the other researchers who have established the method so that it can be easily referenced by readers. It may also be useful to briefly explain what the “low-discrepancy quasi-random sequence” is if relevant to the study.
- Line 48: “After we identify the important tuning targets in the SA stage” – how are those targets usually defined? This hasn’t been explained in the previous paragraph, only that there are different methods for sampling parameters. Please briefly note how that translates to identification of targets (even a sentence should suffice).

- Line 70: “By reusing the 3,840 instances with variations of 11 parameters” – there is no indication of where these numbers are coming from or what they refer to. This should be introduced in the methods section, so wait until that point to elaborate on specifics like this.
- Lines 75-82: please condense into a paragraph rather than bulleted list.
- Lines 77-79: “By reusing the 3,840 sampling instances…” - Again, where does the number of instances come from? And when the authors say the model achieves good accuracy, which variables are they referring to (i.e., the “error within 10%” is an error across which variables?)
- Lines 89-91: “case-specific tuned parameters would further reduce the precipitation error by 15% when compared to a set of unified tuned parameters, and suggest a potential improvement from location-wise parameter tuning in the future.” I did not see the discussion of a case where all cases are combined to find the optimum parameter values. Please elaborate further on that in the results section to support this.
- Lines 98-100: Is it wise to tune for just a single variable (time-mean PRECT)? Realistically, when assessing model performance and tuning accordingly, a *number* of performance metrics need to be accounted for beyond the mean of one variable. Could the authors elaborate on the validity of selecting just one, perhaps?
- Lines 111-114: “In addition, the programs running on Sunway TaihuLight needed to be recompiled due to the adoption of a different archiecture.” - Shouldn’t the model be recompiled at build time? Is this a unique addition to the model, that enables compilation with a non-supported compiler? Is the source code available and going to be included in CESM?
- Lines 184-185: “We set the learning rate…” – could the authors elaborate on if this is the most suitable choice of learning rate/batch size? Were other values tested?
- Figure 5: There’s a relatively wide variability in this across cases, with ARM97 and TOGAII being the least sensitive and GATEIII being very sensitive to the number of parameters to be tuned. Worth elaborating on?
- Lines 225-226: “…although the effect of tuning four parameters was better than tuning three parameters, the advantages of the surrogate model in the parameter tuning process could not be exploited at this point.” I’m unclear why “the advantages of the surrogate model in the parameter tuning process could not be exploited” when using 4 rather than 3 parameters.
- Lines 229-230: “combinations of three parameters lead to the most significance in output” – is this meant to imply the three parameters that drive the most significant *change* in PRECT, or the most significant improvement (assuming those are different, they could be the same, but a big change does not necessarily lead to improvement).
- Lines 255-257: For complete reference, please also explain epsilon and p as they are used in Algorithm 1.
- Line 260: “Meaning average error” – is this meant to be the mean absolute error? Or is this something else?
- Lines 290-292: “We use a total of 7,680 samples, with 1,536 samples for each of the five SCAM case.” - Please elaborate on the reason for choosing this number of samples –

how many values per parameter are enabled by this choice? Is there a clear reason for running 1,536 samples per IOP? This also seems like a detail that should be included in the methods rather than the results

- Lines 298-299: "…although the medium point varies from 5 to 12 mm per day, demonstrating clearly different climate patterns." - How much of this is due to differences in the length of the IOP (perhaps one captured more dry days than another, for example), or an IOP designed to capture shallow vs. deep convection? This may not necessarily be indicative of obviously varying climate patterns.
- Lines 318-319: "The reason for this difference probably comes from a different time of year and the forcing field simulated in these two cases." It would be good to see a more confident assertion here. How different is the time of year assessed, is it substantial enough to cause such a change in sensitivity? How different is the forcing field (and does this hypothesize that it's the large scale T or Q convergence that's responsible)? Is there a difference in the type of convection that occurs as well?
- Line 322-323: "Instead of calling SA methods directly, we use combinatorial analysis of the magnitude of change to determine the effect that these parameter combinations have on the model output." I'm not sure on what this means, please rephrase for clarity?
- Lines 330-332: "For example, increasing tau tends
- to increase total precipitation in GATEIII, while in the other cases it brings the opposite result." - Is there something special about that case that causes the unique signal?
- Lines 353-353: "The impact of such differences is even multiplied" – unclear what is meant by this statement; what is being multiplied here?
- Lines 366-368: "It is easy to see that in the control experiment there were several spikes where the simulated output was significantly higher than the observed values, as was the case in the first four cases. After tuning, these spikes are significantly weakened and the output is much closer to the observed values." It looks like this is really only an issue in the land-based ARM cases; is that true? It looks like the tuning is still unable to match some of the largest rain rates in GATEIII especially but also TOGAII – is there a reason for that?
- Lines 368-369: "This demonstrates the significance of the parameter tuning provided by the workflow for model." Could the authors be more quantitative here? How much is the bias reduced by, for example?
- Lines 399-400: "It is easy to see that the two land convection regions are closer and, accordingly, the three tropical convection aggregation regions are also closer." It looks like the land cases might be fairly different, particularly in terms of optimal pz4 value. A table would make it easier to compare the 'optimal' values, even if they're given by a range of what's marked in red in Figure 15.
- Lines 402-403: "the distribution of the better value points is different for different parameters even for the same parameter space." Should this read that the better values are different for different *cases* within the same parameter space?
- Line 404: "In the other three cases, smaller values of tau lead to better performance." It looks like the optimal value of pz4 occurs at/near the minimum range for both GATEIII

and TWP06 – have the authors tested expanding the lower limit of this variable further to see if this is the optimum value or if it's being cut off?

- Lines 405-406: "On the other hand, it also shows that there are differences in the distribution of parameters that make the results perform better in different types of cases." Rephrase? This sounds like it's saying the same thing as the sentence before it.
- Lines 407-408: "It can be got that the optimal value points for the two land convection cases are close, while the points for the three tropical convection cases are even closer." While the GATEIII and TWP06 cases are very similar, the difference in the TOGA case challenges the notion that tropical convection cases are closer than what we see in the ARM case. Over land, it also looks like the pz4 optimal values are actually rather different; a table would make this argument more convincing and easier to see, potentially. Or may point to the need for a more nuanced statement.
- Lines 416-417: "The difference between the two cases lies mainly in the time, which therefore reflects that there is also a difference in SCAM's simulation performance for different times." I don't see a time-based sensitivity analysis in this; could the authors clarify/elaborate? Is this a seasonal difference? Were other alternative hypotheses explored to explain the difference in ARM95 and ARM97 (different synoptic conditions, etc)?
- Line 419: "the optimal parameter values for each case can be represented by a vector" – I'm not entirely clear on what this vector would look like; it might help the reader to plot said vector on Fig 15.