Replies to Referee #1, GMD-2022-264

Jiaxu Guo on behalf of all authors

April 18, 2023

Thank you very much for your patient and detailed comments on our work[1]. These valuable comments are very helpful for us to improve this paper. After carefully reading all the questions, we have answered each of them and will make appropriate corrections in the revised version of our manuscript.

In this attachment, the red paragraphs represent your comments, and the black paragraphs below are our corresponding replies.

1 Replies to major comments

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.

We apologize for the unclear statements and the confusion that we may have caused. We totally agree that a single SCAM job would take only a few minutes. The platform we use is based on Sunway Processors. For the 5 cases covered in our article, the shortest one took about 10 minutes and the longest one was no more than 20 minutes. Each Sunway processor consists of 4 Core-Groups (CG). Each CG can support a single MPI process. We normally run one SCAM on one CG (note that the Sunway processor is running at a frequency that is roughly one third of an Intel or AMD processor).

The case we refer to here is a workflow of parameter sensitivity analysis and tuning, which consists of 768 SCAM jobs to run. The one hour mentioned here is the time it takes to assign 768 jobs to the job queue, and to collect the results after all jobs are finished. We do experiments when there are enough resources for multiple times, to compute the time on average. We will add these descriptions in the revised manuscript and try to avoid ambiguities.

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.

As explained above, we apologize for the one hour confusion we have made. When we say greater than 1 hour, we do not mean that a case run with a single IOP would take 1 hour, but that a run with 768 different cases would take 1 hour. We record the time in a normal supercomputing environment, so as to demonstrate the related time overhead for scheduling, running the job, as well as collecting the results.

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.

We agree that our work is primarily about making a large-scale parametric analyses possible, which is what you mean by "tractability". Although SCAM itself has a very short run time, the results of the analysis can be obtained faster by further reducing the resource overhead of the experiment in a large-scale experiment. Again, this is an effort to improve the tractability of the analysis. We will adjust the description of this key issue in the revised manuscript.

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.

We will try to avoid redundancy with the background described earlier and elaborate more on the methodology we have used in the revised manuscript.

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 ofparameters? 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.

This number is based on the number of samples from the MOAT and Saltelli sampling methods. In order to reconcile the two sampling methods used in the text, a number was chosen that is large enough and that matches the relationship between the number of samples generated by both methods. 768 is the number of samples per IOP case. We will add this in a revised version of the manuscript.

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.

Here we are mainly describing the environment in which the experiments in this paper were run. This platform is really not strongly correlated with the implementation of the experiments, and their results. We will remove this full section on use of Sunway TaihuLight in the revised version of the manuscript.

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?

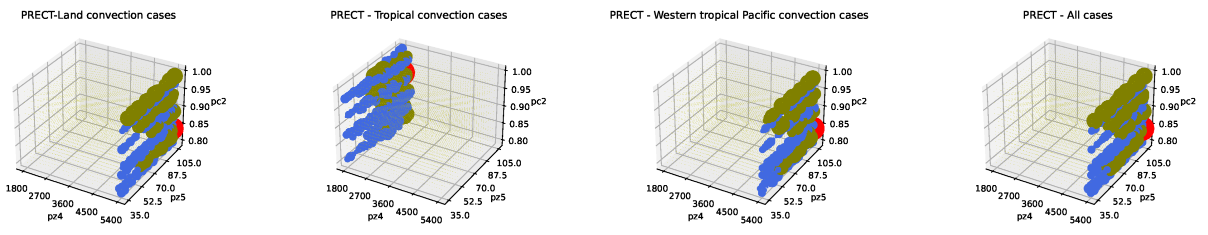
In Section 4.6, we use the optimal solution of each case as a vector, combined with the Pearson correlation coefficient method to calculate the similarity between the individual cases. By using the coefficient as a metric, it is possible to get a more intuitive view of the relationship between these cases. This part of the vectorization analysis is based on Section 4.5. The range of values taken in the Section 4.5 test is the same for all cases. We will add an experimental exploration of all cases, and cases of the same type using the same parameters, as shown in AC-Figure 1.

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?

We are sorry for the misunderstanding. When a draft of this article was written, there was a section 4.7. It was removed in a later edit, but we did not correct the narrative in time.

"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.

This is a copy editing error in the collaboration and should read as follows, "After the scenario where the parameter configuration is the same among cases has been considered, the closeness between the cases could be analyzed." We have also added experiments using the same parameter values for all cases. This is shown in AC-Figure 1. From this we can see the distribution of the output in the parameter space for all cases or cases belonging to the same type, when they take the same parameter values.



(a) (b) (c) (d)

AC-Figure 1: Results of a three-parameter full-space grid search for the multi-objective scenarios using the surrogate model. The points closest to the observed data are shown in red, those ranked 2-64 are shown in olive, and those ranked 65-256 are shown in blue. Where (a) is the scenario of optimizing two land convection cases with the same set of parameters, (b) is the scenario of optimizing three tropical convection cases with the same set of parameters, (c) is the scenario of optimizing two western tropical Pacific cases with the same set of parameters, and (d) is the scenario of optimizing all five cases with the same set of parameters.

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.

The results in our paper are derived from experiments using SCAM to confirm optimal parameter values. In fact, this confirmation is already included in the parameter tuning workflow proposed in this paper. We will describe the experiments we have carried out more explicitly. We will clarify that the final results used for comparison are the result of the SCAM simulation process and that the surrogate models are only used to provide parameter values. This is in line with the ultimate aim of this paper, which is to apply the optimized results to SCAM.

We have also carefully read each of the detailed descriptions you mentioned in relation to the text and captions, including the units consistently on the y-axis.

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?

It is the difference between lowest and highest value of PRECT output. The output value here refers to the average value throughout the simulation, for each case. We will add some necessary information in our revised manuscript.

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.

We are sorry that the mistaken reference has caused confusion. The reference here should be to Equation 1 and not to Table 4 in the original manuscript. We will add more detail here in the revised version of our manuscript. What we are trying to convey here is that, as can be seen by Equation 1, when calculating the effect of the combined parameters on the results, the N_{MPP} increases exponentially as p increases due to the position of p in the exponential. For ease of reading, we have also included Equation 1 in this attachment.

$$N_{MPP} = C \times \binom{M}{p} \times L^{p} \tag{1}$$

Figure 9: Which SA method is being used in this figure? Units on the y-axis would also be helpful.

The single parameter perturbation method is used here, i.e. keeping the other parameter values constant at their default values and tuning only the value of one parameter linearly. To illustrate the problem more clearly, we will also add the units of the y-axis.

Figure 12: Please add units to the y-axis.

We will add units to the y-axis in Figure 12 in the next manuscript submission.

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?

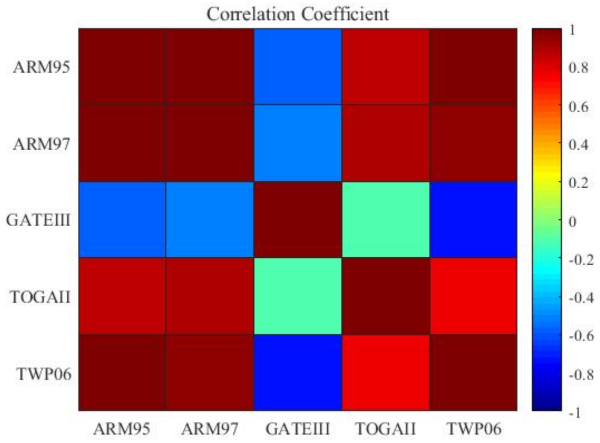
The blue bars indicate the improvement in effectiveness relative to the control experiment when using the traditional SA method combined with a single optimization method. The three bars show the improvement of the three different test approaches compared to the control test. The control test is used as the baseline for the three scenarios here. We will add more appropriate descriptions in the revised manuscript.

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?

The units used are the number of hours it takes to perform a simulation. As NN's improved approach relative to *Original* is mainly reflected in the sensitivity analysis, and this part of the experiment does not involve running more SCAM instances, the change in computing time is not reflected significantly, and therefore the difference in computing time is less reflected. We will refine these descriptions in the revised manuscript.

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?

This might be related to the pre-processing that the vectors undergo before they are involved in the calculation. Given that we have re-trained the model, new results will also be presented in our revised version, as shown in AC-Figure 2.



AC-Figure 2: Correlation of the optimal solutions of the cases in the same parameter space.

2 Replies to 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.

We will add this specification in a revised version of the manuscript.

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.

We will make the appropriate changes in the revised abstract and remove the specific number of instances.

Line 15: Should this read "the effects of global climate change"?

We will correct this in the revised version of the manuscript according to your comment.

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.

We will refine the citations to the references in the revised manuscript and ensure that the links are all accessible. References [2] and [3] will be added to make the descriptions more precise.

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?

The intention here is to illustrate the complexity of CAM and thus set the scene for the introduction of SCAM below. We will revise these descriptions as: "The use of SCAM for large-scale experiments is more practicable due to its advantage of lower requirements for computing resources."

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?

The main purpose here is to highlight the complexity of GCM and thus illustrate where the advantages of SCAM lie. We will rephrase these descriptions in the revised version as :"Considering the uncertainty of schemas, it is more convenient to select models with low computational cost, such as SCAM, to conduct large-scale parameter tuning experiments."

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).

We will correct this issue in the revised version. All notations for the same definition will be unified.

Line 24: "good alternative model" – rephrase for clarity. What's meant by 'good' here – cheaper, more efficient, etc?

Cheaper computational overhead and higher efficiency are both advantages. We will rephrase it to make this more clearly expressed.

Line 25: What is meant by SCAM only needs "one process"?

Since SCAM is a small and fast model, it runs only on one processor. In one simulation of SCAM, only one process is required for each run of one case to complete the computation.

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.

The term 'participants' refers to the independent variables in the problem under study. We will describe this in more concise terms in a revised version of the manuscript.

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.

We will refine these descriptions so that the reader can more easily understand the role that this quasi-random sequence has played in previous studies as well as in the present 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).

Targets, in this context, refer to the parameters (or a combination of them) that are more sensitive to the results obtained during the SA phase of the analysis. Having identified these targets, readers can then know which parameters to tune to have a greater impact on the results, thus making it easier to get better tuning results. We will refine these descriptions as "After we have determined the combination of parameters to be tuned".

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.

We will complete the description of the origin of this sample of figures in the Introduction.

Lines 75-82: please condense into a paragraph rather than bulleted list.

We will condense and refine this part of the text in a revised manuscript.

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?)

For the sample size, we will go into more detail as to why this sample number was selected. The latter achievement refers to the improvement in the model fit to total precipitation, i.e. the variable is PRECT.

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.

In the revised edition we will add clarifications and refine the descriptions so that the reader can understand them more easily. As shown in Figure 1.

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?

PRECT is an output variable included in the IOP files of all five cases covered in this manuscript, and its use as an object of study facilitates cross-sectional comparisons between the cases and the analysis of their relationships.

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?

As we chose Sunway TaihuLight as our experimental platform, we were able to make the code work on this system by attempting to port and compile it. By modifying the source code of the model, the parameters involved in the paper can be tuned via the namelist input file. As a result, there is no longer any need to recompile each time one experiment is carried out, which also makes it much more efficient. The code is currently available.

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?

The values chosen are empirical parameter values of learning rate and batch size that we commonly use for neural network model training. This paper focuses on the feasibility of using neural network methods for large-scale parameter analysis and tuning, and therefore the empirical values were chosen for testing. This set of parameter values is a selection of the better performing values after testing several sets of values. We have conducted an ablation experiment for learning rate and batch size and will detail the process and results of this experiment in a revised manuscript.

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?

What this figure reflects is indeed of some interest. The original meaning of the figure was how much output fluctuation (in terms of average precipitation during the simulation) could be produced for each case when tuning one to four parameters, respectively. The graph does show that for the ARM97 and TOGAII, their precipitation response for four parameters is even smaller than the response of the GATEIII for tuning one parameter. An important reason for this is that these two cases themselves have smaller precipitation values than GATEIII, whereas the figure uses the absolute values of precipitation. This phenomenon does deserve further elaboration and will be discussed in the revised version.

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.

Testing combinations of parameters on surrogate models is also computationally costly. As the size of the test increases exponentially with the number of parameters to be tuned, the computational time will no longer be negligible when tuning four parameters. The results also show that the improvement of maximum tuning effect that can be achieved by tuning four parameters is limited compared to tuning three parameters. Therefore, considering both the computational overhead and the tuning effect, we chose to tune three parameters for the experiments in this paper.

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).

Theoretically, the more parameters that can be tuned in one experiment, the greater the variation in the results that can be brought about. The reason for choosing to tune three parameters here is also to achieve a balance between computational overhead and tuning effect. This allows the sensitive parameters to be tuned while avoiding the non-critical parameters consuming computational resources.

Lines 255-257: For complete reference, please also explain $epsilon$ and p as they are used in Algorithm 1.

We will explain in detail the meaning of these two variables and the role they play in this algorithm. ϵ is the threshold at which the results converge and p is the total number of parameters to be adjusted.

Line 260: "Meaning average error" – is this meant to be the mean absolute error? Or is this something else?

It refers to MAPE (Mean Absolute Percentage Error). In addition, after careful discussion, we will instead use RMSE (Root Mean Square Error) as a measure of error. We will correct it in the revised version.

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.

As with the answer to the question on line 70, since this research involves two methods of generating sample sequences, MOAT and Saltelli, each of which has a different formula for generating the number of samples. Their respective base numbers are different and the choice to generate 1536 samples per case (768 for each of the two methods) was also made in view of the fact that 768 is a appropriate sample size that can be generated by both methods.

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.

Indeed, the five IOPs involved in this study differed in their location, length of capture and time of day. For example, GATEIII has a longer capture time, while TOGAII has a relatively shorter capture time.

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?

Forcing fields are an important cause of the different properties of these IOPs. The difference between the cases is also the focus of our interest and attention. Indeed, the gap between ARM95 and ARM97 indicated by the experimental results in this paper is worthy of attention. By comparing their forcing fields, we find that there is indeed a certain degree of difference between them in T and Q.

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?

This refers to the invocation of the method proposed in this paper for combined parameter analysis, rather than the existing fixed method.

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?

This case does show a unique signal of interest at this point, which may be related to the nature of the GATEIII case itself. A similar conclusion was reached when we explored the parameters with the help of the surrogate model.

Lines 353-353: “The impact of such differences is even multiplied” - unclear what is meant by this statement; what is being multiplied here?

In this sentence, we are trying to express the level of impact by using the word *multiplied*.

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?

This phenomenon was indeed more pronounced in the two land-based ARM cases. On average, our tuning is also effective for the three tropical convection cases. The inability of the tuning to match the total precipitation for GATEIII does exist, as can also be seen from the analysis of the sampling results in Figure 7(c). The failure to cover the total precipitation that matches the observations in the sampling results means that the likelihood of finding the optimal solution through tuning is small. In contrast, this possibility is still present in TOGAII. The existence of this result is justified by the fact that there is a certain margin of error in the simulation of the model itself.

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?

The main purpose here is to highlight the important role that scientific workflow plays in the work of this paper. The methods we present in this paper are organized in the form of a workflow, and the entire reconciliation process is done coherently. We will describe this more quantitatively in the revised version.

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.

We agree that the specific values should be more intuitive and readable for the reader. We will add this to the revised edition.

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?

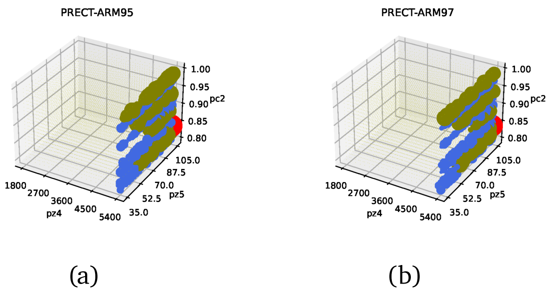
We will correct this in the revised version so that it will be easier for the reader to understand.

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?

Indeed, as can be seen from the figure, the optimized values obtained for both cases do lie at the boundary. Thus the possibility exists that there may be better values outside the bound than inside that bound. However, the values of the parameters cannot be infinitely large or small, and we should also take into account their physical meaning.

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.

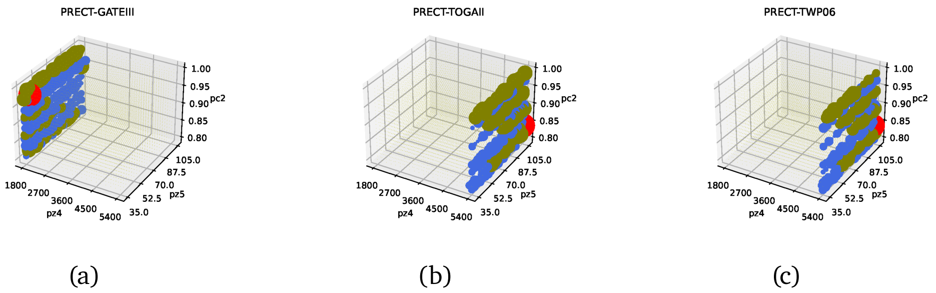
Indeed, as you say, we will use more concise phrases in the revised version: "This reflects the fact that it may be useful and necessary to adopt different parameter configurations for different cases or regions."



(a)

(b)

AC-Figure 3: Results of a three-parameter full-space grid search for ARM95 and ARM97 using the surrogate model. The points closest to the observed data are shown in red, those ranked 2-64 are shown in olive, and those ranked 65-256 are shown in blue. Where, (a) is ARM95 and (b) is ARM97.



(a)

(b)

(c)

AC-Figure 4: Results of a three-parameter full-space grid search for GATEIII, TOGAII and TWP06 using the surrogate model. The points closest to the observed data are shown in red, those ranked 2-64 are shown in olive, and those ranked 65-256 are shown in blue. Where, (a) is GATEIII, (b) is TOGAII and (c) is TWP06.

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.

Indeed, as you have said, further details could be given in terms of specific optimized values. A more detailed elaboration would make our experimental results and views more convincing. At the same time, we conducted another complementary experiment, which was to try to introduce different methods to train the surrogate models for SCAM cases. Through our research, we learned that both XGBoost[4] and ResNet[5] can be used to perform regression tasks and train surrogate models. Here, we will compare the effectiveness of several methods such as LR (Linear Regression), RF (Random Forest), MLP (Multi-Layer Perceptron), XGBoost and ResNet for training surrogate models. The results are shown in AC-Table 1. The RMSE was used to measure the error generated during training. Based on these results, we used ResNet to retrain the surrogate models, and when we used these later trained models to perform a grid search in the same parameter space, we found that TOGAII and TWP06 have more similar distribution patterns in the parameter space, as can be seen from AC-Figure 4. This is also consistent with the above distribution of the two cases in terms of position. This is due to errors in the previous training models, and we would appreciate your prompt correction.

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)?

The term 'time' here refers not to a time-based sensitivity analysis, but to the historical time simulated by these SCAM cases. At the same time, we conducted another complementary experiment, which was to try to introduce different methods to train the surrogate models for SCAM cases. Through our research, we learned that both XGBoost[4] and ResNet[5] can be used to perform regression tasks and train surrogate models. Here, we will compare the effectiveness of several methods such as LR, RF, MLP, XGBoost and ResNet for training surrogate models. The results are shown in AC-Table 1. The RMSE was used to measure the error generated during training. The best performers are shown in bold. Based on these results, we used ResNet to retrain the surrogate models, and when we used these later trained models to perform a grid search in the same parameter space, we found that their distributions were very similar, as can be seen from AC-Figure 3. This is due to errors in the previous training models, and we would appreciate your prompt correction.

AC-Table 1: RMSE of different surrogate models for five cases.

Case	LR	RF	MLP	XGBoost	ResNet
ARM95	0.235	0.197	0.751	0.184	**0.038**
ARM97	0.188	0.158	0.555	0.136	**0.045**
GATEIII	0.646	0.432	1.335	0.538	**0.137**
TOGAII	0.179	0.112	0.223	0.118	**0.041**
TWP06	0.344	0.220	0.594	0.220	**0.040**

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.

The term *vector* here refers to the optimized solution of a set of parameters as a vector. This is because the computation of the Pearson correlation coefficients is done on the basis of vectors. We will refine this description in a revised version.

References

[1] J. Guo, Y. Xu, H. Fu, W. Xue, L. Wang, L. Gan, X. Wu, L. Hu, G. Xu, and X. Che, “A learning-based method for efficient large-scale sensitivity analysis and tuning of single column atmosphere model (scam),” *Geoscientific Model Development Discussions*, vol. 2022, pp. 1–28, 2022. DOI: 10.5194/gmd-2022-264. [Online]. Available: https://gmd.copernicus.org/preprints/gmd-2022-264/.

[2] J. T. Bacmeister, M. F. Wehner, R. B. Neale, A. Gettelman, C. Hannay, P. H. Lauritzen, J. M. Caron, and J. E. Truesdale, “Exploratory high-resolution climate simulations using the community atmosphere model (cam),” *Journal of Climate*, vol. 27, no. 9, pp. 3073–3099, 2014, Cited by: 163; All Open Access, Green Open Access. DOI: 10.1175/JCLI-D-13-00387.1.

[3] J. M. Dennis, J. Edwards, K. J. Evans, O. Guba, P. H. Lauritzen, A. A. Mirin, A. St-Cyr, M. A. Taylor, and P. H. Worley, “Cam-se: A scalable spectral element dynamical core for the community atmosphere model,” *International Journal of High Performance Computing Applications*, vol. 26, no. 1, pp. 74–89, 2012, Cited by: 258. DOI: 10.1177/1094342011428142.

[4] W. XingFen, Y. Xiangbin, and M. Yangchun, “Research on User Consumption Behavior Prediction Based on Improved XGBoost Algorithm,” en, in *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA: IEEE, Dec. 2018, pp. 4169–4175, ISBN: 978-1-5386-5035-6. DOI: 10.1109/BigData.2018.8622235. [Online]. Available: https://ieeexplore.ieee.org/document/8622235/ (visited on 04/03/2023).

[5] L. Shi, C. Copot, and S. Vanlanduit, “Evaluating Dropout Placements in Bayesian Regression Resnet,” en, *Journal of Artificial Intelligence and Soft Computing Research*, vol. 12, no. 1, pp. 61–73, Jan. 2022, ISSN: 2449-6499. DOI: 10.2478/jaiscr-2022-0005. [Online]. Available: https://www.sciendo.com/article/10.2478/jaiscr-2022-0005 (visited on 04/03/2023).