Summary of changes, GMD-2022-264

Jiaxu Guo on behalf of all authors

May 16, 2023

Dear Editor and Referees:

Thank you very much for your patient and detailed comments on our work. These valuable comments are very helpful for us to improve this paper. We have answered each of comments and made appropriate corrections in the revised version of our manuscript.

In this attachment, we will summarize the main modification changes in this paper.

(1) In training the surrogate model, we introduced five different regression methods to train separately to compare their fitting effects. Based on this, we rerun all the experiments. The corresponding figures of experimental results have been also replotted. We also used more quantitative data to show the effectiveness and reliability of our method. The newly trained surrogate models have a better fit and less inconsistency with the actual model.

(2) When training the surrogate model, we also included the samples generated by the Morris sampling method in the dataset, which doubles the size of the original dataset and further improves the fit. RMSE was used as the loss function instead in all experiments.

(3) For multi-site problems, we also carried out optimization for each of the four scenarios. Separately, they are: for ARM95 and ARM97 cases, for three tropical convection cases, for TOGAII/TWP06 (they're close in location) and all five cases, as shown in Figure 14.

(4) In Section 1, we rearranged the word order of the narrative, and elaborate according to the status of pattern research, existing parameter analysis and regulation methods, the challenges faced, and the contributions of this paper. At the same time, we also expanded the scope of literature survey and introduced more literature to further introduce the current research status.

(5) We further clarified the running time of SCAM and try to avoid any ambiguity. Meanwhile, some statements in the manuscript have been revised to make them more transparent.

In the following, we will reply to comments one by one.

Replies to Referee #1, GMD-2022-264

Jiaxu Guo on behalf of all authors

May 16, 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 remove the ambiguous statements and focus on clarifying the problem in terms of the number of computations. The revised text is on Lines 229-241.

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. The revised text is shown on Lines 240-241.

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 have adjusted the description of this key issue in the revised manuscript, as detailed in Line 79-86.

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 have removed this part of the redundant description, as detailed in the revised Section 3.1.

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 have added this in a revised version of the manuscript, as detailed in Section 3.1.

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 have removed this whole subsection in the revised version of the manuscript, as detailed in Section 4.

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?

Section 4.5 mainly looks at a more efficient optimization process and the final result of the individual case optimizations. Section 4.6 focused more on the distribution of results across cases in the same parameter space. We show this rule through 3D figures, which is more beneficial for us to analyze the similarities and differences between these cases. Meanwhile, 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. The range of values taken in the Section 4.6 test is the same for all cases. We redo the experiments again, and we also added the experimental exploration of all cases, and cases of the same type using the same parameters. The content of this subsection has been comprehensively revised, including Figure 13-15.

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. We have corrected the error here.

"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 Figure 14. 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.

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. The experiments in Section 4.5 were performed by finding the optimal solutions using the surrogate model and testing them in the complete SCAM. The resulting error pairs between the surrogate model and SCAM are shown in Table 8 in the revised version. Experiments in Section 4.6 were performed in surrogate mode due to their heavy computational burden. We have revised the above description.

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. Due to the adjustment of the graph order, we have revised on the caption of Figure 6.

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 (4) but 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 (4), 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.

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 have also added the units of the y-axis.

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

We have added units mm/day 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?

We redo the experiment and draw the figure. In the revised version, it is Figure 11. CTL refers to the control experiment using the default value. Baseline refers to an experiment using only one optimization method. Optimized refers to the optimization experiment that combines the surrogate model and grid search.

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 total computational 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.

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 Figure 15.

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 have added 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 have removed the specific number of instances.

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

We have corrected 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] have been 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 have revised 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 :"The use of SCAM for large-scale experiments is more practicable due to its advantage of lower requirements for computing resources."

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 have corrected 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 have rephrased 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 have rephrased it in the revised version of the manuscript, as detailed in Line 28-31.

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.

Thanks for your suggestion, we have revised the description according to the relevance.

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 have refined 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 have corrected this in the revised manuscript.

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

We have condensed 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.

We have included a corresponding discussion in the revised Section 4.6, as shown in Figure 14.

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 added details 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 figure 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.

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 also needs computational resources. 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 have explained 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 instead use RMSE (Root Mean Square Error) as a measure of error.

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. We have made this point in addition in Section 3.1.

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. We have reorganized the language of Section 4.1 (originally Section 4.2).

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?

We are sorry for the mistake. This problem comes from the error generated in our previously trained model. This problem has been resolved after retraining the surrogate model. ARM95 and ARM97 show consistency in all aspects, which is also in line with our expectations.

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. We have reformulated the expression.

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. Different from other cases, this case is a ocean case and is located in the Atlantic Ocean.

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*. We have rephrased it in the revised version.

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 have added quantitative analysis 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 have added Table 9 in 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."

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 Table 6. 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, as can be seen from Figure 14. 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 Table 6. 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 Figure 14. This is due to errors in the previous training models, and we would appreciate your prompt correction.

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 have refined this description in Section 3.5.

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 col-

umn 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).

Replies to Referee #2, GMD-2022-264

Jiaxu Guo on behalf of all authors

May 16, 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 refree comments, and the black paragraphs below are our corresponding replies.

1 Replies to major issues

In this paper, although the authors evaluate the accuracy of the NN model in terms of precipitation, it probably exists the inconsistent between ML and real model, which don't be highlighted in this paper.

We agree that it may be difficult for a surrogate model to be fully consistent with a real model. We use the RMSE in training as a loss function to verify the correctness of the method, that is, whether the parameter tuning of SCAM can be accelerated by training a surrogate model, and to compare different regression methods. We will add more descriptions about the inconsistency to the manuscript from two perspectives: From a model training perspective, the value of loss function indicates the error in the training. However, the process by which we train the model is also the process by which the error is gradually reduced. From a practical perspective, when we use the surrogate model for optimization, the solutions obtained are also validated in the original SCAM case to ensure as much consistency as possible, as shown in Table 6 and 8.

The authors believe that due to the high computational cost of the GCM, the SCAM can be the alternative model for parameter SA and tuning. In reality, the optimal parameters tuned in SCAM could not be suitable for GCM, due to the global regions and more complex large-scale circulation.

We agree that the solution set obtained by parameter tuning on SCAM is not directly applicable to GCM. However, through our attempts at parameter tuning on SCAM cases located in different regions, we can find commonalities and patterns in the parameter response of these cases. We believe such an idea can be applied to the parameterization scheme of the GCM. Our exploration of parameterization solutions using SCAM is mainly from a methodological and ideological point of view. That is, SCAM is a simpler and less costly way to perform numerical simulations. The training of a surrogate model for SCAM is a further extension of this idea.

Overall this manuscript, the organization and writing are not clear and should be well structured. There are a very large number of language errors. The English writing should be greatly improved.

We will reorganize the structure of the article in a revised version to rationalize the logic. We will also work on improving the English writing style to make it more fluent.

The authors separately tune the parameters in SCAM for each site and get the different sensitive parameters and different optimal values. It is difficult to transfer this information to GCM. If the authors can do the multi-objective tuning for these sites with the same parameters, it could be helpful for global model tuning because these SCAM sites indeed represent the different regimes.

We agree that a combined optimization that tries to minimize the differences against observations across all five cases would be more meaningful. Such tests and results were included in a previous version of the draft. However, we removed the results at certain stage for a more focused description in the result section.

As suggested by the reviewers, we have carried out a careful analysis and done corresponding experiments. After retraining the surrogate model for each of the five cases by using ResNet, we also carried out multi-objective optimization for each of the four scenarios. Separately, they are: for ARM95 and ARM97 cases, for three tropical convection cases, for TOGAII/TWP06 (they're close in location) and all five cases, as shown in Figure 14.

Combined with results shown in Figure 13 of our grid search for the five cases individually, it can be seen that cases located at the same or similar locations really have similar distributions of precipitation output in response to parameters. Therefore, this kind of joint optimization based on the multi-objective idea is of general interest.

For the workflow, there should be a "metrics" component because it is very important for tuning. No matter SCAM or GCM, the tuning metrics could be the cost function between model simulations and observations. The different designs could affect the optimization. In terms of the metrics, it could consider the 1) different statistic errors between simulation and observation, such as RMSE, performance score like Yang et al. (2013), 2) one objective or multiple objectives, and how to deal with the multiple objectives.

I agree that "metrics" are important factors to consider in the workflow. We will use RMSE as the metric of error in the revised version.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{1}$$

After using RMSE as the metric for training the model and parameter optimization, the corresponding results were recalculated. In particular, the losses during the training of the model are shown in Table 6. The 3D parameter space responses for the five cases themselves and for the four multi-objective joint conditioning scenarios are shown in Figure 13 to 14.

Regarding multi-objective tuning, we also designed corresponding experiments when we originally wrote this paper. Four scenarios are included, separately, they are: combined tuning for ARM95 and ARM97 cases, combined tuning for three tropical convection cases, combined tuning for TOGAII/TWP06 and all five cases. From results we can see that the parameter responses for ARM95 and ARM97 have a similar distribution and thus the results of the multi-objective optimisation for both of them reflect this. The same trend is reflected similarly in the TOGAII and TWP06.

Line 35: The statement that the Morris SA method cannot get the interactive sensitivity could be wrong. Aurally, the standard deviation of MOAT samplings can stand for the interactive effect of one parameter with others (Morris, 1991).

We have refined the description of the Morris SA method in the revised manuscript so as not to introduce ambiguity in Line 37-39: "Morris SA can give the individual sensitivity of each parameter, including their interaction sensitivity. However, this is not intuitive enough if the user wants to know directly from a combined perspective which set of parameters has the most significant effect on the results."

Line 45: as the part of introduction, the authors should explain the challenge of the SA methods, why you choose Morris and Sobol, the computational cost issue, surrogate problems using machine learning. If there are previous works, what's your contribution?

We will complete this part based on a further full investigation. In the revised version, we will give a more detailed explanation. Both Morris and Sobol are typical SA methods that have a wide range of applications in many fields. As there are already proven application examples, it makes sense to conduct experiments based on the above methods. In addition to this, we have introduced several new SA methods that have been proposed in recent years. Although SCAM, as a single-column model, already consumes less computational resources than GCM, the computational resources of a system are not infinite. To further explore their parametric features, a study using surrogate model is necessary. Surrogate models [2] can significantly reduce the computation time of individual tasks, thus making it possible to scale up experiments.

Our contribution lies in the fact that we have trained the surrogate models on SCAM using a regression-based approach, and with the help of the surrogate models we have conducted parameter sensitivity tests for combinations, as well as tuning for the most significant parameter combinations.

Line 55: the authors should do comprehensive literature research, even for GCM, there are a large number of work for tuning, such as Yang et al. (2013) and Zou et al. (2014). In addition, the NN surrogate model is used to tune as well. But the authors don't introduce the previous work and challenge in terms of this issue. The introduce section should be more clear.

We will complete the introduction section by conducting a more detailed literature survey of the work related to the content of this paper. This also includes the literature you mention, such as [3] and [4]. Yang et al. [3] analysed the sensitivity of nine parameters in the ZM deep convection scenario for CAM5 and used the SSAA (Simulated stochastic approximation annealing) method to optimize the precipitation performance in different regions by zoning. Zou et al. [4] conducted a sensitivity analysis for seven parameters in the MIT-Emanuel cumulus parameterization scheme in RegCM3. The precipitation optimization process for the CORDEX East Asia domain was carried out using the MVFSA (Multiple very fast simulated annealing) method.

The current challenges lie in the following areas. (1)No similar experiments have been conducted on SCAM. The short run time of the SCAM makes it easier to obtain more samples in a short period of time and thus scale up the experiments. (2)The usual SA methods will give the sensitivity of the individual parameters. But when we look at a set of parameters, is the best combination of N parameters the top N sensitivity of a single parameter? This is a question worth exploring. (3) As various neural network methods are applied to the field of regression, more appropriate regression methods should be applied to the parametric study of Earth system models. Unlike the various public data sets commonly used, ESM-based experiments will also further enrich the practical implications of research in the field of regression analysis.

In our manuscript, we innovatively use a neural network-based agent model for parameter tuning of different cases in SCAM. For each case studied in the paper, the surrogate models are trained separately based on different methods, and the best performing model is selected by their training errors (RMSE).

Line 75, Acutely, there are existing SA and tuning workflow used in climate models, such as PSUADE and DAKOTA, the authors don't compare their workflow to these packages. It's not new for the community.

This is indeed something we need to improve further in the literature survey. We have done a survey of the packages such as PSUADE, DAKOTA and STATA etc., including some studies based on their work in different fields. These packages can indeed implement the functionality of SA and tuning. We also compared the workflow proposed in this paper with the above software packages. From a method perspective, we added the comparison of the new SA methods in recent years, such as RBD-FAST, Delta and HDMR. The above approaches haven't been fully supported by all the packages above. For using neural networks to train surrogate models, DAKOTA currently only supports neural networks with a single hidden layer. Meanwhile, the proposed method uses more types of neural networks and supports the adaptive selection of the best performing network to train the surrogate model.

Line 88: The authors don't mention the 30% improvement for tuning error and computational cost. How do they come from?

The tuning error is compared with control experiments. This is the result of a simulation using the default parameter values of SCAM. The improvement in computational overhead comes from comparing it with a traditional optimization workflow. In the original optimization workflow, the original SCAM is invoked for each calculation of the objective function. In the approach proposed in this paper, the objective function is replaced with a surrogate model. This allows the overall workflow execution time to be compressed, thus the computational overhead could be reduced.

Table 2 is wrong. Each IOP file includes many variables, not just these four variables. Therefore, the statement that you choose precipitation is wrong.

We are sorry for the misunderstanding. We agree that there are many variables in the IOP file for each case, not just these four, as can be seen from Table 2. We include Table 2 in our manuscript mainly to illustrate that the variables contained in different IOP files are different. We will add this in the revised manuscript.

Line 120: This issue could not be a significant challenge. Some simple scripts can collect the data.

Yes, scripts are able to do that. Of course, we will also achieve efficient management of experimental data in our method. Due to the amount of data generated during this stage, better data management is necessary. We have adopt a more reasonable description in the revised manuscript.

We have revised these statements in Line 127 as follows: "In this step, we also integrate the collection and processing script for the post-sampling results. As the output of SCAM is stored in binary files in NetCDF format, the precipitation variables we want to study need to be extracted from a large number of output files in order to proceed to the next step. This will further accelerate the degree of automation of scientific workflows and thus accelerate the conduct of research in this area of the earth system models."

The statement about Morris is wrong, see 3.

We plan to correct this issue in Line 38 as follows, "Morris SA can give the individual sensitivity of each parameter, including their interaction sensitivity. However, this is not intuitive enough if the user wants to know directly from a combined perspective which set of parameters has the most significant effect on the results."

In section 3.1: sampling is not equal to SA. In this section, the authors introduce the SA methods. You should consider change the structure or change the title.

We agree that sampling can't be confused with SA. We will refine the article structure of this section in a revised version to make the expression easier to understand. The sampling method and the SA method will be split into two subsections to be described.

Subsection 3.1, entitled "Sampling of SCAM", will be devoted to the sampling methods covered in this paper and the rationale for their selection.

Subsection 3.3, entitled "Sensitivity analysis for a single parameter and combinations of parameters (enabled by the NN-based surrogate model)", will provide a more detailed description of the SA methods covered in this paper and the rationale for their selection. At the end of this subsection, our proposed combined SA methods for supplementary validation will be introduced.

Line 177: it could be better to compare NN with other surrogate models, such as xgboost, ResNet.

Yes, the properties of the various surrogate models are something we all care about. Through our research, we learned that both XGBoost[5] and ResNet[6] 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 Table 6. Since ResNet has better performance in model training, we will choose ResNet as the network for training the surrogate model. Subsequent experiments will also be supplemented in the revised paper.

Line 184: The 768 samples seem not enough for training NN, do the authors evaluate the performance of NN? In Figure 4, how do you define the accuracy?

Yes, we have evaluated the performance of NN and measured its ability on regression. After our analysis of the sampling results, we found that 768 samples can already cover the range of values of the parameters and output variables. Therefore, the number of parameters selected is sufficient to obtain a good training effect. The results of the training process and the experimental results also support our conclusion. For the second point, The R^2 score is used as the accuracy rate in Figure 4. It is defined as

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \tag{2}$$

where $\hat{y}$ is the predicted value and $\bar{y}$ is the mean of the test set. In addition to this, RMSE was also used as an important metric to measure the loss during training.

Line 185: Do the authors do the hyper-parameter tuning for NN?

In the initial version of our manuscript, our aim was mainly to verify the feasibility of this approach, so we used a set of empirical parameters to train the NN. Now we will conduct hyperparameter tuning trials in NN training to make the proposed method more solid. We have refined these in Table 5.

Line 195: Due to the uncertainties of each method, ensemble can't guarantee to reduce the error.

We agree that there are errors in out initial manuscript. We feel sorry for the term "integrate" in the original paper may bring some ambiguity. We've used several SA methods to make a more intuitive side-by-side comparison, which allows us to find the best method for each case. It's not necessarily about reducing the error.

Line 210: Equation (1), the left hand of this equation is not the number of processes. It should be the number of simulations. The number of simulations could depend on different sampling method. For Morris, it could not require such number of samples. It is not clear for this description.

Since SCAM is a single-process task, one simulation is equivalent to one process. Of course, as you said, it is necessary to distinguish between simulation and process more accurately. For Morris, 768 was used as the sample number to keep the sample number of various sampling methods consistent. We will add the above contents to refine these issues in a revised version of the manuscript.

For a SCAM simulation, it usually requires 10-20 minutes, why do you require more than one hour?

We agree that for a single SCAM job it only takes one process to run, and it does not take very long for a single SCAM to run. The one hour mentioned here is mainly the time it takes to complete a whole workflow of parameter analysis and tuning. This includes multiple iterations and delays in queuing batch jobs. We will refine these descriptions above in the revised manuscript and try to avoid ambiguities.

It is confusing that you can re-use the sampling from SA to train the surrogate model for tuning, but in section 3.3, you mention that the surrogate model can be also used in SA?

We agree that the trained surrogate model can be also used in SA. Since we want to find the most sensitive set of parameters in combination, this process requires a larger experiment scale. Therefore, to perform large-scale parametric analysis experiments faster, we use surrogate models to speed up the process.

Line 233: how do you get the conclusion? It is not convinced.

Here, the running time is compared to one full simulation of the original SCAM. Here we are trying to make two points. On the one hand, the surrogate model can simulate the output of SCAM more accurately. This makes it an effective alternative to SCAM in terms of parameter response. On the other hand, the execution time of the surrogate model is very short, which saves a lot of time for parameter analysis and optimization tuning. Combining these two reasons above, the use of a surrogate model for parameter tuning of SCAM is a very reasonable strategy. We have also revised the description in this subsection.

It is very confusing for section 3.4 and is difficult to follow your idea. You could consider to re-organize this section.

We're sorry for not being able to make it easy for you to follow our idea. We have reorganized this subsection to make it easier to understand. In this subsection, we focus on an enhanced SCAM parameter tuning process. There are two main contributions. On the one hand, we use the trained surrogate model as the objective function for optimization, which allows the optimization time to be compressed considerably. On the other hand, we propose an enhanced parameter optimization process based on grid search. We use grid search to reduce the search range in the optimization process and thus obtain better results in fewer iterations. In the revised manuscript we will elaborate further on these two points above.

Line 270: how many samples do you have for the correlation? Do you do the p-value test?

Here, we carried out correlation analysis on the respective optimal solutions of the five

SCAM cases. Therefore, a total of five vectors were used to calculate the correlation For the second point, now we have added tests on p-values in the revised manuscript.

Figure 7: it is difficult to evaluate the tuning performance in Figure 7. It could be better to use metrics like Yang et al. (2013).

What we want to express in Figure 7 is the statistics of the sampling results. The phrase "when the parameters are tuned" refers to the fact that the parameters are given different values during the sampling stage, not the final tuning stage. We're sorry for this ambiguity and will respect your suggestions to revise the description appropriately. When revising, we fully refer to the relevant statements of [3] in order to make the results more clear and easy to understand. The results after reanalysis are shown in Figure 3.

Figure 4, Figure 5 are the results but appear in section 3 (methods). It could be reorganized.

The results described in these two figures belong to the pilot test, which is to prove the rationality of our experimental ideas. We respect these suggestions and provide appropriate revisions to make the organization of our paper more rigorous. The relevant statements have already been described in Section 4.

Line 313: pz2 (c0_ocn) should be high influence on the ocean case. But in Figure 8, it doesn't have the high effect on PRECT at TWP. Could you explain the reason?

It can be seen from Figure 8 that no matter which sensitivity analysis method is used, pz2 (c0_ocn) has a certain influence on the results. It is a deep convection parameter related to the ocean-land intersection, but this does not mean that it must be the one that has the most influence on TWP. The influence of the ocean has been demonstrated here.

Line 317: the reason for the different between ARM 95 and ARM 97 is not convinced.

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. We retrained the surrogate model using ResNet and used RMSE as the metric of error. We found that the response of the precipitation output to the parameters in the new surrogate model was very similar for both cases which can be shown in Figure 13-14. This suggests that in the original version, the difference between the two examples is due to the error in the model. Your advice on the choice of the neural network would also be much appreciated here.

Line 345: are the 16 iterations enough for convergence? Why don't use the general optimization, such as PSO, GA that you mentioned in the introduction section?

In our experiment, 16 rounds of iteration can already make the results converge, as shown in Figure 9. So the selection of the number of iterations is sufficient. WOA[7] is chosen here mainly because it is an optimization algorithm proposed later. Theoretically speaking, in the optimization stage of the workflow proposed in this paper, no matter what kind of optimization algorithm is selected, the purpose of optimization can be achieved, including PSO, GA, etc.

2 Replies to minor issues

In Fig. 1, is there an arrow pointing from "SA methods" to "testing of combinations"?

There isn't an arrow here because only the trained surrogate models are used for testing combinations of parameters.

Caption in Figure 1: The sentence "SCAM launcher, the data collector and the jobs therein represent the batch execution of the SCAM algorithm" should be rephrased. What are the "jobs" and "batch execution"? It is not clear.

There isn't an arrow here because only the method of the surrogate model trained using neural networks will be used for parameter combination SA. Here, a job refers to a single simulation of SCAM. A batch is a collection of jobs that are submitted to the computing queue at once. We will elaborate more in detail in the revised manuscript.

Line 25: should explain "ne30"

This is a description of the spectral element method grids[8]. It refers to a model grid with a $ne30np4$ spectral element (approximately 1-degree) atmosphere and land grids. $ne[X]np[Y]$ are cubed sphere resolutions where X and Y are integers. The short name generally is $ne[X]$. The corresponding descriptions have been reorganized.

Line 38: the reference of Sobol method is wrong, pls use the original paper.

We will correct it [9] in the revised manuscript.

Line 43: The QMC and LHC are sampling methods, not SA methods.

We will correct this in the revised manuscript. Sampling methods and SA methods will be more clearly distinguished to avoid confusion.

In Table 3, how do you select these parameters? And how do you define the range of each individual?

These parameters were chosen mainly from reference [10]. The parameters range from 50% to 150% of the default value.

Line 94: there should be a reference for SCAM5.

We have added reference [11] in the revised manuscript to refine the introduction of SCAM5.

Line 102: all sites belong to ARM.

We have corrected this issue in the revised manuscript. We will also carry out a more detailed study later.

Line 108: "in the code" change to "in the model"

We have corrected it in the revised manuscript.

Line 116: "is an important issue" change to "are important issues"

We have corrected it in the revised manuscript.

Line 130: It is only suit for SCAM. For GCM, it is impossible.

Indeed, as you say, it is not practical to run large batches of GCM jobs even on HPC. We have refined this description in the revised manuscript.

Line 165: please explain the "second-order sensitivity"

It refers to the mutual influence between two parameters. We have refined this exposition in the revised manuscript.

Table 4: the reference of Sobol is wrong.

We have corrected this reference [12] in the revised manuscript.

Line 174: please consider the correct position of this sentence.

We have corrected it in the revised manuscript.

Line 193: It is not clear for "not a direct evaluation."

We have added some necessary information and details in our revised manuscript. What we want to express here is the following: using the above methods, the set of M parameters that most influence the result may be difficult to obtain directly. The method proposed in this paper just fills this gap.

Line 195: add "have" before "its"

We have corrected it in the revised manuscript.

Line 215: how do you get the number of multiple thousand?

Combining Equation (1) with the practical problem studied in this paper, we can see that when $C = 5, p = 3$ and $L = 10$, 5,000 simulations are needed even if only one parameter combination is considered. This will be astronomical when more combinations are considered. We have also refined the above in the revised manuscript.

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] K. Cheng, Z. Lu, C. Ling, and S. Zhou, "Surrogate-assisted global sensitivity analysis: An overview," en, *Structural and Multidisciplinary Optimization*, vol. 61, no. 3, pp. 1187–1213, Mar. 2020, ISSN: 1615-147X, 1615-1488. DOI: 10.1007/s00158-019-02413-5. [Online]. Available: http://link.springer.com/10.1007/s00158-019-02413-5 (visited on 04/03/2023).

[3] B. Yang, Y. Qian, G. Lin, L. R. Leung, P. J. Rasch, G. J. Zhang, S. A. Mcfarlane, C. Zhao, Y. Zhang, and H. Wang, "Uncertainty quantification and parameter tuning in the cam5 zhang-mcfarlane convection scheme and impact of improved convection on the global circulation and climate," *Journal of Geophysical Research Atmospheres*, vol. 118, no. 2, pp. 395–415, 2013.

[4] L. Zou, Y. Qian, T. Zhou, and B. Yang, "Parameter tuning and calibration of regcm3 with mit–emanuel cumulus parameterization scheme over cordex east asia domain," *Journal of Climate*, vol. 27, no. 20, pp. 7687–7701, 2014. DOI: https://doi.org/10.1175/JCLI-D-14-00229.1. [Online]. Available: https://journals.ametsoc.org/view/journals/clim/27/20/jcli-d-14-00229.1.xml.

[5] 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).

[6] 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).

[7] S. Mirjalili and A. Lewis, "The Whale Optimization Algorithm," *Advances in Engineering Software*, vol. 95, pp. 51–67, 2016, ISSN: 18735339. DOI: 10.1016/j.advengsoft.2016.01.008.

[8] (). "Model grids – cime master documentation," [Online]. Available: http://esmci.github.io/cime/versions/master/html/users%5C_guide/grids.html.

[9] I. M. Sobol, "Sensitivity analysis for non-linear mathematical models," *Mathematical modelling and computational experiment*, vol. 1, pp. 407–414, 1993.

[10] Y. Qian, H. Yan, Z. Hou, G. Johannesson, S. Klein, D. Lucas, R. Neale, P. Rasch, L. Swiler, and J. Tannahill, "Parametric sensitivity analysis of precipitation at global and local scales in the community atmosphere model cam5," *Journal of Advances in Modeling Earth Systems*, vol. 7, no. 2, 2015.

[11] P. A. Bogenschutz, A. Gettelman, H. Morrison, V. E. Larson, D. P. Schanen, N. R. Meyer, and C. Craig, "Unified parameterization of the planetary boundary layer and shallow convection with a higher-order turbulence closure in the community atmosphere model: Single-column experiments," *Geoscientific Model Development*, vol. 5, no. 6, pp. 1407–1423, 2012. DOI: 10.5194/gmd-5-1407-2012. [Online]. Available: https://gmd.copernicus.org/articles/5/1407/2012/.

[12] I. Sobol, "Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates," *Mathematics and Computers in Simulation*, vol. 55, no. 1, pp. 271–280, 2001, The Second IMACS Seminar on Monte Carlo Methods, ISSN: 0378-4754. DOI: https://doi.org/10.1016/S0378-4754(00)00270-6.