A Summary of the Major Changes

Dear Editors and Referees,

We are grateful for your thoughtful and detailed review of our manuscript. Your comments and suggestions have been invaluable in helping us improve the quality and clarity of our work. We appreciate the time and effort you have invested in providing such constructive feedback. We have carefully considered each of your points and have addressed them as follows.

- We have redefined the abstract, fundamental motivations, innovative aspects, and primary contributions of this paper.

- We have conducted more comprehensive research into this field and improved the references accordingly.

- We have restructured the framework of this paper and streamlined the language, aiming for a more concise and straightforward presentation.

- We have expanded the exploration to include additional variables such as humidity, temperature, and cloud.

- We have corrected some minor details in the original content of the manuscript.

Best regards,

All authors

Replies to Referee #1, GMD-2022-264

Jiaxu Guo on behalf of all authors

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.

Comment 1: Line 17: Bacmeister et al. (2014) is not the correct citation for CESM, please refer to recent documentation of the full model rather than CAM alone. Similarly, the reference for CAM in the next sentence should be updated; the current citation (Dennis et al.2012) refers to the development of a spectral element dynamical core; more recent and broader citations are available.

Thanks for your comments. We have corrected the above-mentioned issues in the revised manuscript[2, 3].

Comment 2:Line 22: "SCAM. . . [is] a cheaper and more efficient alternative model for the purpose of tuning physics parameters." This discussion of SCAM and its use deserves more nuance. The authors here seem to suggest that parameter tuning can be done purely in SCAM and that SCAM is a suitable surrogate for running the full global model. While SCAM is a useful tool for initial testing of parameterizations (and perhaps tuning), it is not in fact a full substitute for running CESM globally.

Thanks for your comments. We apologize for any confusion caused. The main objective of this paper is to use machine learning methods to construct a surrogate model for SCAM and thereby assist in its parameter tuning. It is not suggested here that SCAM itself is a surrogate model for a global model, although some conclusions may provide insights for our study of global models. We have also removed ambiguous sentences and corrected corresponding descriptions in the text.

Comment 3: Line 29: ". . . has the most significant effect on results. and the Sobol method that uses the. . . " Sentence fragment re: Sobol method, perhaps leb over from previous edits?

Thanks for your comments. We apologize for the typo here, and we have removed the duplicated sentence fragment.

Comment 4: Lines 67-69: These toolkits are not well introduced; if they can implement SA and tuning, what exactly is the downfall of using them? Please elaborate on what niche this is filling. The authors state that the use of more parameters and cases increase the exploration space exponentially and make the task "almost an impossible job," but don't show clear evidence of that. Line 76 states that the benefit of the approach is the addition of "new SA methods in recent years." – which methods? Is that addition the only benefit (in which case it doesn't seem that the expansion to more cases/parameters is nearly impossible in those packages, as stated)?

Thanks for your comments. We apologize for any confusion in this section. Currently, there are various sensitivity analysis (SA) methods, and this paper introduces 5 different SA methods on an equal footing. Additionally, it compares these methods with our own multi-parameter joint perturbation method based on machine learning.

Regarding the second question, despite SCAM having relatively low computational cost, in cases where a large number of executions are required, the computational cost is still non-negligible. Therefore, the introduction of surrogate models can further reduce the computational cost in the experimental process, expediting the experimental progress.

Comment 5: The reference to Pathak et al. (2022) is missing a DOI and is incomplete.

Thanks for your comments. We have improved the expression of this reference[4].

Comment 6: Line 106: In general, the selection of PRECT is perhaps reasonable given its importance societally and the focus on convective parameterizations. Yet it seems like Pathak et al. assessed the response of a number of climate fields to their perturbations, which more closely mimics the process of model tuning. Is it not worthwhile to expand to more indicators of model performance? Near surface T/Q or surface fluxes, for instance, seem to be available for most IOPs.

Thanks for your comments. We highly appreciate your point of view, and in this revision, we have included an exploration of temperature, humidity, and cloud as you suggested.

Comment 7: Line 110: The bounds for parameters are selected to be 50% and 150% of the values; are there not more firm limits based in the literature for what a realistic range may be?

Thanks for your comments. We determined the range for parameter tuning based on general physical principles. The primary consideration for making this choice is the concern that exceeding certain limits may affect computational stability.

Comment 8: Lines 112-114: please indicate which fields you've now made tunable parameters (vs. which ones already were).

Thanks for your comments. In the original code, there are adjustable parameters such as zmconv_c0_lnd, zmconv_c0_ocn, zmconv_ke, cldfrc_rhminh, cldfrc_rhminl, and others, while the rest are not adjustable. After modifying the code, we made all parameters studied in this paper adjustable. We have also provided corresponding descriptions in the text.

Comment 9: Lines 116-118: As before, I'm still not entirely clear on what the authors mean by

saying the programs needed to be recompiled and that the recompilation has enabled a larger number of concurrent instances on the Sunway supercomputer. I believe that this is referring to the fact that all parameters in table 3 are now namelist options, thus the model does not need to be re-built each time and can instead be cloned from a single, pre-built base case whose namelist can be modified. That does cut down on run time, but doesn't seem related to the programs running on Sunway needing recompilation due to the supercomputer planorm. Suggest clarifying that this is simply the shift of hard-coded to namelist defined variables.

Thanks for your constructive comments. Frankly speaking, the platform on which it runs does not ultimately affect the conclusions of this paper. Therefore, in this revision, we no longer emphasize issues related to the platform and compilation, as they are not the focal points we need to address.

Comment 10: Line 174: "we decide another S=768, where NMorris is 64 and NSaltelli is 32." What's the sensitivity to those choices? Does 64 and 32 show some convergence to a solution?

Thanks for your comments. We apologize for the ambiguity introduced here. The intended meaning is that, with this configuration, the same total number of samples can be obtained under both sampling methods. Since these two N values are not critical factors in the experiment and to avoid any misunderstanding, we have removed the corresponding description in the revised version.

Comment 11: Line 187: "Compared to other networks, ResNet has the advantage of using less pooling." What does this mean? Should elaborate on what pooling is first if this is an important point to make. If the following explanation is intended to elaborate on why ResNet is preferred, perhaps rephrase for clarity.

Thanks for your comments. We apologize for the confusion caused here. The main advantages of ResNet (Residual Network) lie in addressing the issues of vanishing gradients and exploding gradients during the training of deep neural networks. In this paragraph, we have also reorganized the corresponding language.

Comment 12: Line 190: "ResNet18. . . " similar question to above – what's the sensitivity to number of layers? Is 18 enough (and how have the authors determined that)?

Thanks for your comments. In terms of network depth, taking ResNet50 as an example, due to its deeper architecture, ResNet50 has significantly more parameters than ResNet18. While having more parameters can generally make the network more flexible, it may also lead to overfitting. Considering the specific application context and the size of the dataset in this study, ResNet18 is deemed sufficient to meet the requirements of the research.

Comment 13: Lines 208-209: "As the sensitivity values calculated by different methods are of different orders of magnitude. . . " – was this expected, and arises as a by-product of using different SA methods? I.e., are the units not all in the form of some response driven by a change in the parameter?

Thanks for your comments. We apologize for the confusion caused here. What we intended

to convey is how to identify the most impactful combination of parameters in the scenario of combined parameter perturbation. To make the expression more concise, we have removed the potentially ambiguous sentences in the revised version.

Comment 14: Lines 232-233: "This way, we can determine the number of parameters in the combination, taking into account the tuning effect and the amount of computation." – Not clear what's meant here. What is meant by 'number of parameters in the combination,' isn't that set? How is the 'tuning effect' taken into account, as this seems like a tool that aids in tuning rather than the tuning have an effect on the process.

Thanks for your comments. This refers to whether we should adjust 3 parameters or just 1 or 2 parameters. How to choose which parameters to adjust? Indeed, our method can assist in tuning, but selecting the most appropriate parameters also plays a crucial role in effectively enhancing the tuning effect.

Comment 15: Line 234: "lead to the most significance in output" – should this read the most significant improvement in output, or the largest increase in accuracy? Could be interpreted as the biggest change in output as written currently.

Thanks for your comments. Here, indeed, it refers to "the biggest change in output." We have improved the corresponding descriptions accordingly.

Comment 16: Line 292: "...the distribution of pz4(tau)." –if zmconv_tau in figure, suggest consistent naming conventions. In general, it would be more useful to use the zmconv_tau naming convention as in Fig 3 for Fig. 5 as well, to make it easier for the reader to quickly scan through.

Thanks for your comments. We have revised the corresponding descriptions in the above figures.

Comment 17: Figs 3 and 4 – using consistent colors in Figure 4 to represent the cases shown in Fig. 3 would increase the readability.

Thanks for your comments. We have readjusted the color scheme of the figures to maintain consistency and enhance readability.

Comment 18: Fig 4 – since the main point is the increase in accuracy of PRECT rather than just the change, perhaps there's some way to indicate the change in precipitation bias rather than the overall percent change? It's possible this belongs more in Figure 3 (i.e., gevng a sense of how much the bias has improved given each parameter, rather than just which results are "better" than the control).

Thanks for your comments. Your suggestion is excellent. Following this line of thought, we reanalyzed the samples and improved the corresponding descriptions accordingly.

Comment 19: Line 300: "Different from other cases, this case is a ocean case and is located in the Atlantic Ocean" – This seems to suggest that GATEIII is the only ocean case; does TOGAII not also target ocean? If this isn't the only ocean case, what about it being in the Atlantic makes it particularly unique?

Thanks for your comments. We apologize for the confusion caused by our wording. Indeed, GATEIII does exhibit uniqueness in parameter response, partly due to its classification as an ocean case—although it is not the only case located in the ocean. Another possible contributing factor may be its unique geographical location, which warrants further investigation.

Comment 20: Table 5 –doesn't seem to add much to the paper; could this not just be said in the text?

Thanks for your comments. We have now described the corresponding content in the main text.

Comment 21: Lines 309-310: "It can be seen that ResNet has the best performance..." – The RMSE is indeed the lowest, and not just by a lisle but by quite a margin! Was this fully expected a priori? Is this consistent with any other literature?

Thanks for your comments. This is indeed a noteworthy issue. Therefore, we conducted additional experiments and presented the most recent results in the revised paper.

Comment 22: Figure 5: How are the sensitivity results normalized?

Thanks for your comments. Due to the different computational processes of various Sensitivity Analysis (SA) methods, the sensitivity values obtained are not necessarily on the same scale. To facilitate a straightforward horizontal comparison, we normalized the values using the min-max scaling method.

Comment 23: Lines 322-323: "differences between the individual parameters are not as wide as in the other cases" – not sure what's meant here; that the range in parameter values is smaller? The ranges should be consistent across cases, yes?

Thanks for your comments. We apologize for any confusion caused by our wording. What we intended to convey here is the magnitude of changes in the results. Indeed, as you mentioned, the range of variation in input parameters is entirely consistent across all cases. This ensures a consistent basis for horizontal comparison.

Comment 24: Line 324: "This is also consistent with its position" – by position, is the intent to say that the geographic location is more comparable? Please consider rephrasing for clarity

Thanks for your comments. We apologize for any misunderstanding caused by our wording. We have rephrased this to better convey the intended meaning. Here, we aim to express that cases located at different sites do indeed have different sensitive parameters and responses.

Comment 25: Section 4.3: The discussion of sensitivity to SA method should be expanded and clarified. It is stated that the SA results are shown, but which column is this? The Morris test seems significantly more sensitive to parameter variations than the other tests; are the others better suited for one precipitation regime vs. another? Or how is the different sequence of samples coming into play here?

Thanks for your comments. We apologize for any confusion caused. The six columns here

represent the results of different Sensitivity Analysis (SA) methods, each providing distinct analytical conclusions. Since each SA method operates based on different principles, variations among their results are normal. With normalization applied, the Morris SA method's results, which you mentioned, reflect that the sensitivity differences between parameters are relatively smaller compared to other methods, and this is understandable. Our objective is to identify parameters that are more sensitive to the results (in a relative sense).

As the Morris SA method has specific requirements for the sample sequence, we constructed a sequence tailored to it. Except for the method proposed in this paper (the rightmost column), the other four methods use sequences constructed by the Saltelli method. The sample size for each method is the same.

Comment 26: Lines 357-358: "However, the computation amount increases exponentially during the test." But isn't the benefit of using a surrogate model that you can tune more parameters at a time than using the traditional approach? If all tests for 3 parameters can be conducted in less than a minute, why not tune 4?

Thanks for your comments. When the number of parameters simultaneously adjusted increases, the required computational complexity also grows exponentially. In our practical observations, we found that the computational time of the surrogate model is not negligible when simultaneously adjusting four parameters. Considering the results of the sensitivity analysis presented earlier, the parameter ranked fourth in sensitivity has a relatively minor impact on the outcomes. Therefore, we believe that adjusting three parameters is the most suitable choice.

Comment 27: Figure 7: It's unclear what the max/min stars represent or add to the plot; the min stars also aren't visible from what. I'm unsure why the y-axis is positive only; do none of the parameter combinations lead to a decrease in precipitation?

Thanks for your comments. Here, we are describing the range of changes in the output variable when different parameter combinations are adjusted. Since the y-axis values represent the range of output changes, they are non-negative. We apologize for any confusion caused in this context. After a thorough review of the entire document, we have decided to remove this figure to prevent any misunderstanding.

Comment 28: Line 366: "meaning average error" – should this now be RMSE?

Thanks for your comments. We apologize for the typo that occurred here, and we have corrected the description in this section.

Comment 29: Table 8 is somewhat hard to interpret. Is "error" here referring to the difference in SCAM and the surrogate model? Or is it related to either's agreement with observations? Is SCAM doing a better job simulating PRECT at all locations vs. the surrogate model (given the lower RMSE values) – that doesn't seem to be reflected in the text.

Thanks for your comments. The purpose of this table is to demonstrate that the surrogate

model fits well with the original SCAM, making it suitable for carrying out parameter tuning experiments. We apologize for any confusion caused in this regard and have provided a more concise description of this meaning in this section.

Comment 30: Figure 10: Please clarify if the "model simulation output" is from SCAM or the surrogate model. It also makes more sense to plot the x-axis in time coordinates rather than timestep number.

Thanks for your comments. This refers to the numerical outputs from the SCAM simulation. We sincerely appreciate your suggestion and proceeded to redraw this figure.

Comment 31: Lines 377-378: "The tuning of the SCAM parameters is quite productive on the time scale." – not clear what this means with the current wording. What time scale is being assessed? Does 'productive' mean increased accuracy?

Thanks for your comments. What we intend to convey here is that, after applying the method proposed in this paper, the time spent on parameter tuning for SCAM has been reduced, indicating an improvement in efficiency. We apologize for any confusion caused by our previous explanation, and we have revised this description in the newest version.

Comment 32: Line 381: "Although still below the observed level at about 1300 steps of TOGAII, improvement is also reflected." – I'm not sure why this timestep/case is singled out here, it seems like there are plenty of other examples of this. Similarly, the point is made that in TWP06 there are instances where control is less than observed; again, that's present in all cases, yes?

Thanks for your comments. We apologize for the confusion caused by our wording. Here, our main emphasis is to highlight that, after optimization, the output for each case is closer to the observed values compared to the default parameters, thereby demonstrating the effectiveness of our approach. We have corrected the description in the revised version for clarity.

Comment 33: Section 4.5: I'm not clear why the baseline and optimized models should have any different computational cost, unless baseline refers to SCAM and optimized to the surrogate model. If that's the case, that should be stated much more clearly and oben. More generally, when considering overall model performance, the same model structure should be used. So of course use the surrogate model to find the optimum parameters, but to assess differences in simulation accuracy, plug those values back into SCAM. If not comparing direct SCAM cases, the accuracy gains aren't very understandable.

Thanks for your comments. This section aims to emphasize that, with the use of a surrogate model, the time required for parameter tuning and testing can be significantly reduced. We apologize for any confusion caused, and it is clarified that the ultimate optimization results obtained are indeed plugged into SCAM for execution. We have rephrased this portion in the revised version for better clarity.

Comment 34: Lines 388-389: "The use of NN trained surrogate models for parameter tuning can further save computational resource overhead and, in terms of results, can meet or exceed traditional optimization methods in most cases." – this may be overstated. In terms of saving computational

resources, this is not illustrated by Figure 12. Figure 12 includes time in the job queue for the SCAM case; adding a bar to each case for the computational time excluding queue time would be needed to support this statement.

Thanks for your comments. Indeed, as you mentioned, the queuing time during job execution on the cluster should be excluded. We will use a more reasonable approach to articulate the advantages of the proposed method in terms of computational overhead in this paper.

Comment 35: Lines 391-392: "The model can get an enhancement in performance from 6.4%-24.4% in precipitation output," – again, I think the important part is to plug this back in to SCAM using the parameters determined by the surrogate model. Unless the suggestion is to use the surrogate model for all simulations (rather than for finding optimal parameters and model tuning), this is a critical step.

Thanks for your comments. We apologize for the confusion. Our experimental procedure involved taking the obtained parameter values and plugging them back in SCAM. We have corrected the corresponding description in the manuscript.

Comment 36: Table 9: A better description is needed; "multi" cases undefined here.

Thanks for your comments. The intention here is to aggregate similar cases into the same category in an attempt to have them represent a larger region of parameter responses.

Comment 37: Section 4.6: An important point is lacking from the discussion, which is how well does the model perform (RMSE per site) when using the values found to optimize performance across all 5 sites (multi(d))? The potential benefit of regionally-varying parameters is noted, but it's not illustrated how important is it to use potentially regionally refined values vs. ones that are optimized for global performance.

Thanks for your comments. After careful consideration, we acknowledge that aggregating these combinations simplistically to represent the global scenario may be unfair, despite their individual representativeness within certain ranges. Therefore, in this revision, we have abandoned this attempt and focused more on each case, as well as the commonalities and individualities among them.

Comment 38: Related to above, and though this might be outside what the authors' computational resources allow – the most convincing argument for the study would be to apply the tuned parameters from the multi(d) case to a global run. That would illustrate true benefit from using a surrogate SCAM model to find optimal parameter values; without that, the potential impact of the paper is severely more limited.

Thanks for your comments. We highly agree with your perspective. If the methods proposed in this paper can bring about better performance improvements for the global model, we believe the significance of this paper will be further highlighted. As you mentioned, considering the resources currently available to us, conducting identical experiments on a global scale still poses significant challenges. The focus of this paper lies in leveraging machine learning methods and surrogate models for parameter optimization, as well as exploring the differences

in parameter responses across different regions. We will continue to approach this field with great enthusiasm and look forward to conducting more exploration in the future, expanding and validating its applicability further, given the conditions allow.

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. W. Hurrell, M. M. Holland, P. R. Gent, S. Ghan, J. E. Kay, P. J. Kushner, J.-F. Lamarque, W. G. Large, D. Lawrence, K. Lindsay, W. H. Lipscomb, M. C. Long, N. Mahowald, D. R. Marsh, R. B. Neale, P. Rasch, S. Vavrus, M. Vertenstein, D. Bader, W. D. Collins, J. J. Hack, J. Kiehl, and S. Marshall, "The community earth system model: A framework for collaborative research," *Bulletin of the American Meteorological Society*, vol. 94, no. 9, pp. 1339–1360, 2013. DOI: 10.1175/BAMS-D-12-00121.1. [Online]. Available: https://journals.ametsoc.org/view/journals/bams/94/9/bams-d-12-00121.1.xml.

[3] R. B. Neale, C.-C. Chen, A. Gettelman, P. H. Lauritzen, S. Park, D. L. Williamson, A. J. Conley, R. Garcia, D. Kinnison, J.-F. Lamarque, *et al.*, "Description of the ncar community atmosphere model (cam 5.0)," *NCAR Tech. Note NCAR/TN-486+ STR*, vol. 1, no. 1, pp. 1–12, 2010.

[4] R. Pathak, H. P. Dasari, S. El Mohtar, A. C. Subramanian, S. Sahany, S. K. Mishra, O. Knio, and I. Hoteit, "Uncertainty quantification and bayesian inference of cloud parameterization in the ncar single column community atmosphere model (scam6)," *Frontiers in Climate*, vol. 3, 2021, ISSN: 2624-9553. DOI: 10.3389/fclim.2021.670740. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fclim.2021.670740.

Replies to Referee #2, GMD-2022-264

Jiaxu Guo on behalf of all authors

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 refree comments, the blue paragraphs represent comments added by the refree in this round, and the black paragraphs below are our corresponding replies.

Comment 1: 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.

The propose of this manuscript provides a method for efficient tuning of SCAM. This comment question the challenge that the offline surrogate method could not guarantee the optimal solutions in surrogate model can transfer to the real model. The authors try to explain it by the high accuracy of the surrogate model. However, the surrogate models are difficult to learn the optimal solutions because they can't be sampled in the training data. Therefore, the efficiency of this method could be affected. The authors do not statement this point and do not give a solution about this issue.

Thanks for your comments. We fully agree with your point that the optimal solution within the range may not necessarily be present in the samples. This is precisely why we employ machine learning methods to construct a surrogate model, as it adds a more realistic dimension. By learning from the samples, the trained model can capture the relationship between parameter variations and output responses, facilitating quicker identification of the potential optimal solution range. Validating whether a solution obtained from the surrogate model is optimal involves inserting it into SCAM and comparing it with known outputs.

Comment 2: 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.

I do not figure out a clear path from tuning in SCAM to GCM. The authors claim that "tuning on SCAM cases located in different regions, we can find commonalities and patterns in the parameter response of these cases". However, there are several challenges. The current SCAM can only support

several sites and cannot represent the global catachrestic. Second, there are different optimal parameter values at different sites. Even for each site, there are different local optimal solutions. It is difficult to find the so-called commonalities for these optimal parameters. The forcings are different between SCAM and GCM, optimal parameters could not achieve good performance in GCM.

Thanks for your comments. We apologize for any misunderstanding. Indeed, as you pointed out, the optimal solution from SCAM may not directly apply to GCM to achieve good results due to differences in forcings. However, through the exploration of parameter tuning in SCAM presented in this paper, we can distill valuable methodologies. On one hand, by exploring the simultaneous tuning of three parameters, we achieve improvements in model output performance compared to tuning 1 to 2 parameters. On the other hand, building different surrogate models for different cases reflects that the response to parameters varies across regions. This has positive implications for future parameterization schemes. For example, we can establish distinct models for different regions in GCM, exploring the use of parameters that are most suitable for each region.

Comment 4: 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.

For figure 13-14, are the 'ranked 1' points local optimal or global optimal? Different regions (ad) have the different optimal points. It's difficult to determine the optimal point in one region (like a) achieves the optimal solution in the other regions. It's also difficult to find a unified optimal solution for GCM, although it could be helpful to design a new parameterization with different parameter values at different regions.

Thanks for your comments. The points labeled as Rank 1 represent global optima. Indeed, as you mentioned, each region has its own optimal points, and a solution optimal in one region may not necessarily be optimal in other regions. The main significance of Figures 13-14 lies in the exploration of the three-dimensional space formed by the three parameters. It verifies the feasibility of jointly tuning these three parameters. Within the given range of values for the three parameters, we can find a set of more suitable values to improve model performance. This methodology also provides insights for exploring parameter tuning in GCM.

Comment 5: 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.

Does the metric function only consider the precipitation? However, it could consider more variables in GCM. It could make sense using time series RMSE in SCAM. In GCM, mean state, spatial correlation, spatial standard deviation, spatial RMSE should be involved in the metrics. The simple metrics in SCAM also pose the challenge for GCM.

Thanks for your comments. We highly appreciate your point of view, and in this revision, we have included an exploration of temperature, humidity, and cloud as you suggested. Indeed, as you mentioned, there are many metrics worth considering. If we delve further into exploration within GCM in the future, it would be reasonable to judiciously leverage these metrics to enhance optimization outcomes.

Comment 6: 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).

"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." The sentence is confusing. The mean represents the main effect, and the standard deviation represents the interactive effect.

Thanks for your comments. We apologize for any confusion caused by our previous statement. We have corrected the description in the revised version.

Comment 7: 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?

In the revised introduction section, the authors explain why they use Morris and Sobol with the sentence "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 seems this work use them because the previous work used them. This expression is not serious. There are a lot of sensitive analysis methods. The authors should carefully state the advantage of these two methods. The authors also claim the reason of using surrogate that the SCAM requires not very cheap computational cost. However, it only takes several optimal iterations.

Thanks for your comments. We apologize for any confusion in this section. Currently, there are various sensitivity analysis (SA) methods, and this paper introduces 5 different SA methods on an equal footing. Additionally, it compares these methods with our own multi-parameter joint perturbation method based on machine learning.

Regarding the second question, despite SCAM having relatively low computational cost, in cases where a large number of executions are required, the computational cost is still non-negligible. Therefore, the introduction of surrogate models can further reduce the computational cost in the experimental process, expediting the experimental progress.

Comment 8: 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.

I think the explanation of innovation of this manuscript does not make sense. The motivation of this manuscript tried to tune SCAM and transfer to GCM. Since these methods have been applied in

GCM, it is confusing that the authors claim they are not used in SCAM. For the 2nd point, the previous works get the sensitive parameters by perturbing multiple parameters. Then the sensitive parameters can be tuned using optimization algorithms. It is not clear for your point.

Thanks for your comments. We apologize for any confusion. As mentioned in the previous response, the exploration of SCAM does not necessarily involve directly transferring parameter values to GCM. Instead, the process aims to identify patterns and improve the methodology for our research. There may be complex relationships among various parameters, similar to the three-body problem, where adjusting three or more parameters has a collective effect greater than adjusting 1 or 2 parameters. This paper combines machine learning methods with parameter tuning, accelerating the parameter tuning process.

Comment 9: 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.

The explanation is not convinced. The authors do not analyze the advantage and disadvantage of these existing framework. Why do not you improve the existing framework? The references of PSUADE and DAKOTA are wrong.

Thanks for your comments. PSUADE and DAKOTA do have their own advantages, but they primarily focus on traditional numerical and statistical methods, and their deployment is not conducive to running on updated clusters. The proposed approach in this paper introduces a machine learning-based surrogate model construction method. It allows for further exploration of perturbations in multi-parameter combinations, making it flexible and easy to deploy.

We apologize for the shortcomings in the previous literature review and have corrected the references for PSUADE[2] and DAKOTA[3].

Comment 11: Table 2 is wrong. Each IOP file includes many variables, not just these four variables. Therefore, the statement that you choose precipitation is wrong.

Although table 2 lists many variables, this manuscript only consider precipitation in the metrics. It should be considered more variables in GCM.

Thanks for your comments. We highly appreciate your point of view, as answered in the previous question, in this revision, we have included an exploration of temperature(T850), humidity(Q850), and cloud(CLDTOT) as you suggested.

Comment 16: 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?

The authors do not give any result to prove that the 768 samples are sufficient. The offline surrogate model methods are difficult to achieve high accuracy at optimal local regions. The tuning efficiency is easily affected by the local samples.

Thanks for your comments. The appropriate training sample size depends on the specific

machine learning task, model complexity, and the available data. Generally, a larger sample size allows the model to better learn the features of the data and improve generalization. However, an excessive number of samples may lead to overfitting, especially when the model is relatively simple or the data contains noise.

In the specific context of this paper, a sample size of 768 has proven effective in fitting the parameter responses, considering the complexity of the given scenario. Additionally, 768 is a common multiple of the sample sizes generated under the two sampling methods. Regarding the issue of local samples, we have optimized the workflow of the entire method. The solutions obtained from the surrogate model are validated in SCAM, and the validation results are used to enhance the training process. This refinement contributes to further improving the model's fitting capabilities.

Comment 27: 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?

Unfortunately, the author is not familiar with the parameters. C0_ocn is parameter of autoconversion coefficient over ocean in ZM deep convection scheme. It is not the ocean-land interscection.

Thanks for your comments. We apologize for any misunderstanding caused here, and we have corrected the wording in this section in the current revision.

Comment 29: 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?

The statement is very misleading. Any optimization algorithm cannot converge in 16 iterations! This comparison is not serious and it requires at least several hundred iterations.

Thanks for your comments. We apologize for any confusion caused here. In the revised version, we have removed the wording that could lead to ambiguity in this context.

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] C. Tong, "Problem solving environment for uncertainty analysis and design exploration," in *Handbook of Uncertainty Quantification*. Cham: Springer International Publishing, 2016, pp. 1–37, ISBN: 978-3-319-11259-6. DOI: 10.1007/978-3-319-11259-6_53-1. [Online]. Available: https://doi.org/10.1007/978-3-319-11259-6_53-1.

[3] K. R. Dalbey, M. S. Eldred, G. Geraci, J. D. Jakeman, K. A. Maupin, J. A. Monschke, D. T. Seidl, A. Tran, F. Menhorn, and X. Zeng, "Dakota, a multilevel parallel object-oriented

framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: Version 6.16 theory manual," Jan. 2021. DOI: 10.2172/1868423. [Online]. Available: https://www.osti.gov/biblio/1868423.