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.

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.

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 (a-d) 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.

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.

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.

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.

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.

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.

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.

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.

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.

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.