Dear editors and reviewers,

First of all, thank you very much for your time and reviewing our manuscript, we also very appreciate the helpful comments and kind suggestions from the anonymous reviewers. Per the reviewers' comments (in black font), we have revised our manuscript accordingly and made point-to-point responses (in blue font) to all the comments and concerns. Below are our detailed responses.

Reviewer #1

General Comments:

This manuscript is a nice but limited analysis using an automated method for parameter tuning with a global atmospheric model. It is well written and interesting, but limited in scope, and not that well related to previous work. I'm not sure of the statistics, and the dependence of the improvement on (A) the trajectory used and (B) on the specific comparison data sets. I think the paper may be suitable for publication in ACP with some significant revisions along the lines noted below.

Reply: We thank the reviewer very much to take time to review our manuscript. The constructive comments are highly appreciated.

General

1. The paper needs more discussion of previous work. There are several studies cited such as Qian et al 2015 and Zhao et al 2013, Yang et al 2013, that have tried similar approaches with CAM5. Are results consistent, especially with parameters overlapping between many of these studies? This should be mentioned in the discussion.

Reply: Thanks for your suggestion. Qian et al. (2015) and Zhao et al. (2013) investigate the parameter sensitivity related to cloud physics, convection, aerosols and cloud microphysics in CAM5 using the generalized linear model. Yang et al. (2013) tunes serval parameters in Zhang-McFarlane convection scheme using the simulated stochastic approximation annealing method. The previous work requires to run the long-term simulations for sampling and optimization iterations so that the GCM attains physically robust and meaningful signals. In consequence, the whole procedure requires huge computational cost. Our study presents an economic method of automatic tuning by using short-term hindcasts and verifies whether the optimal signal can transfer into the long-term simulation. In this study, the tuning parameters are selected based on the CAM5 sensitivity results of Zhang et al. (2015), Qian et al. (2015) and Zhao et al. 2013. We have added these texts to the introduction section.

Reference:

1. Yang, B., Qian, Y., Lin, G., Leung, L. R., et al.: 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, 118, 395–415, 2013.
2. Zhao, C., Liu, X., Qian, Y., et al.: A sensitivity study of radiative fluxes at the top of atmosphere to cloud-microphysics and aerosol parameters in the community atmosphere model CAM5. Atmospheric Chemistry and Physics, 13(21), 10969, 2013.

3. Zhang, T., Li, L., Lin, Y., et al.: An automatic and effective parameter optimization method for model tuning, Geoscientific Model Development, 8, 3579–3591, 2015.
4. Qian, Y., Yan, H., Hou, Z., et al.: Parametric sensitivity analysis of precipitation at global and local scales in the Community Atmosphere Model CAM5. Journal of Advances in Modeling Earth Systems, 7(2), 382-411, 2015.

2. I am not sure what the statistics are, but how significant is a 10% improvement?

Reply: The significance of improvement depends on the baseline performance under the default parameter values, the structure of the tuning metrics and the optimization algorithm. Firstly, if the default parameter values are well tuned, the optimization magnitude remains little room to improve. We show about 10% improvement in our pre-defined metric for CAM5 that is already a well-calibrated model. Secondly, if the tuning metrics contains lots of model output variables, the tuning complexity and difficulty will increase because the multiple variables have mutual restraint and dependence relationship. Zhang et al. (2015) achieved 9% improvement in terms of comprehensive metrics including 16 variables when tunes a well-calibrated GAMIL2 atmospheric model. On the contrary, Yang et al. (2013) got 31.2% improvement for the annual mean of the convective precipitation when they only tune the single variable. However, the total precipitation got worse than the default simulation. Thirdly, Xu et al. (2018) tuned the parameters for the global soil carbon and got 12% improvement using the global optimization algorithms, but the computational cost of those algorithms is 50~60 times to the algorithm used in this study. Finally, Flato et al. (2013) showed that the change in most of the variables between CCSM4 and CESM (CAM5) kept within the 10% range in term of RMSE. We had added text to illustrate this point.

Reference:

1. Xu, H., Zhang, T., Luo, Y., et al.: Parameter calibration in global soil carbon models using surrogate-based optimization. Geoscientific Model Development, 11(7), 3027-3044, 2018.
2. Flato, G., Marotzke, J., Abiodun, B., et al.: Evaluation of climate models, 2013.

3. I am curious since you state the method is dependent on the initial choice of parameters whether you get the same results if you start with a different set of parameters. Another way of saying this is: how do you know you have a true minimum and not a local minimum in the difference/error metrics?

Reply: Sorry for confusion here. The optimization algorithm used in this study is an improved downhill simplex method. Zhang et al. (2015) shows this algorithm can find a good local minimum solution based on the better choice of the initial parameter values. It cannot guarantee to get the true minimum result. Those global optimization algorithms that can find the true minimum solution always require extreme amount computational cost, such as CMA-ES, EGO and Genetic Algorithm. Zhang et al. (2015) showed the improved downhill simplex method outperformed the global optimization algorithms with the limited optimal iterations.

For clarification, we have added these texts in the revised manuscript in section 3.

4. Most of the improvement is in the LW Cloud Forcing. This is probably not surprising from the choice of parameters, as noted. Please discuss how the choice of parameters might influence the results. Also, I am concerned that the ISCCP data set is a bit old: what happens if you (A) compare the result optimized on ISCCP with CERES-EBAF LWCF (And SWCF), or (B) use CERES-EBAF in the estimation. I am very curious and this paper would be of more utility if it showed sensitivities to observational data sets used, and whether they are of sufficient quality that training against different data sets yields different answers.

Reply: Zhang et al. (2015) showed that LWCF is highly sensitive to changes in the CAPE consumption time scale (zmconv_tau) and the minimum rh for high stable clouds (cldfrc_rhminh). Yang et al. (2013) also indicated the tau was sensitive for LWCF. The autoconversion efficiency of cloud water to precipitation (zmconv_c0_lnd and zmconv_c0_ocn) and the cloud ice sedimentation velocity (cldsed_ai) were found to be sensitive for LWCF in Qian et al. (2016). That is to say, all tuning parameters in this study are very sensitive for LWCF, resulting in this field to have the most improvement. These texts are added to section 4.1 on describing the results in Table 1.

The following Table compares the AMIP estimation results using ISCCP or CERES-EBAF in terms of the percentage bias of LWCF and SWCF. The parameters are optimized on ISCCP in the short-term hindcast simulation. The percentage bias is defined as:

$$bias\|\%\| = |\frac{\overline{EXP}(\overline{CNTL}) - \overline{OBS}}{\overline{OBS}}| \times 100\%$$

Observations	Default	Tuned
SW Cloud Forcing (ISCCP)	3.603	5.116
LW Cloud Forcing (ISCCP)	17.607	8.643
SW Cloud Forcing (CERES-EBAF)	4.477	2.836
LW Cloud Forcing (CERES-EBAF)	14.528	5.223

LWCF is accomplished by a significant improvement with respect to both ISCCP and CERES-EBAF. SWCF with respect to ISCCP sees an increased error. It is consistent with the tuning result in the short-term hindcasts, shown in Table 3 in the manuscript. On the contrary, SWCF on CERES-EBAF achieve a better result than the default simulation. At the very least, this suggests using ISCCP for the optimization is acceptable. We decide not to make change in this aspect, considering this paper is more about the auto-tuning approach than the result itself.

Reference:

1. Qian, Y., Wan, H., Rasch, P., et al.: Parametric sensitivity in ACME-V1 atmosphere model revealed by short Perturbed Parameters Ensemble (PPE) simulations,

https://climatemodeling.science.energy.gov/sites/default/files/presentations/Qian-ShortSimulation-2016SpringMeeting-ACME_Poster.pdf

Specific Comments

5. Page 2, L28: if you tweak parameters in it you need to mention the stratiform cloud microphysics. You are mentioning the ice sedimentation velocity, so please describe the microphysics.

Reply: Thanks for your suggestion. We have added it in the revised manuscript.

6. Page 3, L4: verification of hire tuned model.

Reply: Corrected

7. Page 3, L8: hindcasts of July 2009 are used

Reply: Corrected

8. Page 3, L9: All of the 3 day simulations

Reply: Corrected

9. Page 3, L13: hindcasts are typically from a single day, using 3 days errors may be evolving, and that may bias results? It looks like the biases are systematically evolving from figure 1 between day 2 and 3. How does that affect the results?

Reply: Using 3 days is a trade-off between having the short-term hindcast results consistent with AMIP simulations on the metrics involved and keeping the computational cost in check. Xie et al. (2012) indicated that the ensemble of the 5th day simulation has a good similarity with the AMIP results. Our Figure 1 shows that the results from the interval-Day3 is similar to that based on the single 5th day. while at just 1/5 of the computational cost. We consider using interval-3Day format is reasonable in capturing model biases due to physics in AMIP simulations. Calibration with the interval-Day3 construct has a consistent improvement with the AMIP simulation as shown in Table 3.

Reference:

1. Xie, S., Ma, H., Boyle, J., et al.: On the correspondence between short-and long-time-scale systematic errors in CAM4/CAM5 for the year of tropical convection, Journal of Climate, 25, 7937–7955, 2012.

10. Page 3, L18: The NCEP/NCAR analysis is an older product with known biases. Are you sure these do not affect results? Maybe also add a more modern reanalysis?

Reply: Thank you for the suggestion. The results also hold when comparing with the ERA-Interim. Table 4 in the manuscript includes comparison of AMIP results with ERA-Interim in terms of the percentage bias of vertical mean of the relative humidity and the temperature. The error of relative humidity increases slightly, and the error of temperature decrease after tuning. These relative changes are the same in Table 3 for Q850 and T850 that are with respect to NCEP reanalysis. We decided to stay with using the NCEP reanalysis for the optimization.

11. Page 3, L21: while testing how the model...
Reply: Corrected

12. Page 4, L14: missing a description of how this relates to previous CAM5 tuning exercises.
Reply: Thanks for your suggestion. We have added a sentence in the manuscript in Section 3.1. "The uncertainty ranges of the parameters in this study are based on Covey et al. (2013) and previous CAM5 tuning exercises (Yang et. al., 2013; Qian et. al., 2015)."

13. Page 5, L12: how sensitive are results to choice of metric, weighting, etc? Why does a zonal mean reduce biases in short term forecasts? I don't follow the logic. If the biases are not random, they will not average out.
Reply: We agree that zonal mean certainly does not reduce hindcast bias itself. Use of zonal mean is more for the features to focus on during optimization and to have the optimization be effective. There always exists small spatial displacement between the short-term forecasts and observations. This mismatch at grid scale will commonly lead to large errors in grid-by-grid MSE that could marginalize the responses due to parameter changes. Using zonally averaged fields allows the optimization to focus on the effective responses at global scale. The original text is clear in illustrating the impact of using grid-scale errors in metric calculation. In the revision at the end of section 3.2, we added text to further clarify the purpose of using zonal mean.

14. Page 6, L2: how do you know there is not a local minimum using this method? Have you tried approaching from different initial locations in the parameter space?
Reply: Please see reply 3 for the first question. We tested the random initial parameter values in Zhang et. al. (2015). The well-selected initial locations achieved the better optimization than the random initial values using less computational cost.

15. Page 6, L24: I think you should speculate here or in the summary about what these parameters do. Using sedimentation velocity increases ice mass which increases the LWCF cloud forcing.
Reply: Thanks for your suggestion. How the parameter change affects the simulation is discussed in Section 4.2. We added a sentence in the manuscript in Line 273. "How the parameter change affects the simulation is discussed in Section 4.2"

16. Page 6, L26: define moderate and noteworthy. Do you mean significant in some way? These are vague concepts. If you start from a different place do you get a different answer?
Reply: The sentence has been revised as "This is relatively a significant reduction". We also added additional text to illustrate it. Please also see the reply 2 regarding the significance of the 10% improvement.

To answer your question regarding starting from a different place, we think very likely it would lead to a different answer. In this study, we use the improved downhill simplex algorithm. One key aspect of the improved process is to select good initial values for this algorithm. Choices of the initial values play a very important role in the optimization performance and computational cost.

17. Page 6, L32: what is the observational dataset used to compare the LWCF? Is it isccp? If so, CAM5 was probably tuned on CERES, a different data set. How much would that matter? What if you used a different data set?
Reply: Yes, it is ISCCP. Please see the reply 4 that we use the CERES EBAF to evaluate the optimization results which are tuned on ISCCP. The conclusion is the same when using the ISCCP to evaluate, that LWCF has a big reduction in term of MSE.

18. Page 8, L25: a warmer atmosphere
Reply: Corrected

19. Page 9, L9: how much does the choice of variables matter? If you chose liquid micro-physical parameters, would you see more optimization in SWCF?
Reply: Thanks for your suggestion. We will add the liquid micro-physical parameters in our further work to improve SWCF. The different choice of variables in metrics will likely lead to different optimization results in certain quantities.

20. Page 9, L14: is this improvement significant in any sense?
Reply: Please see the reply 2.

21. Page 9, L19: is the improvement in the broader metrics smaller?
Reply: Yes, we show another metrics that compute the percentage bias of the ten fields between the default/optimized model and the reference observations, which is computed based on 2-dimensional monthly mean fields as the follows:

$$bias\|\%\| = |\frac{\overline{EXP}(\overline{CNTL}) - \overline{OBS}}{\overline{OBS}}| \times 100\%$$

Some reference observations are different from those used in tuning process. The two metrics have the consistent optimal direction for the five fields used in tuning process.

22. Page 19, Figure 4: if this is cloudsat only it is not an appropriate comparison as cloud- sat misses high thin cloud the model represents. Please explain the figure In more detail.
Reply: Sorry for the confusion here. The data used are actually composite of collocated CloudSat and CALIPSO measurements. We have revised it in Figure 4.

23. Page 20, Figure 5: probably not the right data set. Are results the same with CERES EBAF.
Reply: We elected to use ISCCP data for the comparison, as it is the same data used in the optimization. The relative improvements in mean biases and RMSE are similar when comparing with CERES-EBAF.

24. Specifically: if you optimize on one data set, and compare against another, do you still see improvement?

Reply: We do see that, please also see the replies 4, 10 and 23.

Reviewer #2:

** Summary ** This is a nice study that investigates the possibility of automatic tuning, by using a much more efficient set of successive short simulations with CAM. The outcome is that the model can be improved for an aggregate set of objective parameters through an objective minimum finding algorithm. There are however discussions that could be included to clarify the outcomes of the analysis and give greater insight into the performance of the algorithm, for all the individual variables.

Reply: We gratefully thanked the reviewer for the time and very helpful comments and suggestions. We have revised our manuscript accordingly.

1. It's pretty clear to order 0 that LWCF is the only optimized variable by this process. Do the authors know why this is?

Reply: The tuning parameters selected, as indicated in Yang et al. (2013), Zhang et al. (2015) and Qian et al. (2016), were all very sensitive for LWCF so that this field had the most of the improvement. We have added related text in the 3rd paragraph of section 4.1.

Reference:

1. Yang, B., Qian, Y., Lin, G., Leung, L. R., et al.: 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, 118, 395–415, 2013.
2. Zhang, T., Li, L., Lin, Y., et al.: An automatic and effective parameter optimization method for model tuning, Geoscientific Model Development, 8, 3579–3591, 2015.
3. Qian, Y., Wan, H., Rasch, P., et al.: Parametric sensitivity in ACME-V1 atmosphere model revealed by short Perturbed Parameters Ensemble (PPE) simulations, https://climatemodeling.science.energy.gov/sites/default/files/presentations/Qian-ShortSimulation-2016SpringMeeting-ACME_Poster.pdf

2. LWCF has always been a poorly performing variable in CAM and the main reason is that the clear-sky component of the LW is too strong, but their combination give about the right total LW at the top of atmosphere. This is the often the reason that LWCF is chose to be weaker than observed in order to get an accurate TOA flux. So does LWCF improve in CAM5 at the expense of the total LW?

Reply: Unfortunately, as you pointed out, the improvement does come at the expense of the total LW. But the consistency between hindcast tuning results and that from AMIP simulations suggest the tuning strategy remains effective.

3. Models are more often tuned (at least by humans!) from an energy budget balance and coupled system perspective. How does this technique inform this standard approach, and what would need to be added to make it appropriate? Will coupled work as well as AMIP type?

Reply: We expect that the relative responses to parameter changes can work as well with coupled mode as with the AMIP type, because the tuning strategy focuses on the responses of atmospheric fast physical processes, which makes the use of short-term hindcasts appropriate.

However, this tuning strategy cannot be directly applied to the coupled system because of much longer time scales needed for the surface systems to respond. Given the computational efficiency of the tuning approach described in this work, we think it can be used to provide guidance on which parameters to select as well as the direction to vary the parameters for hand tuning the model towards a certain target.

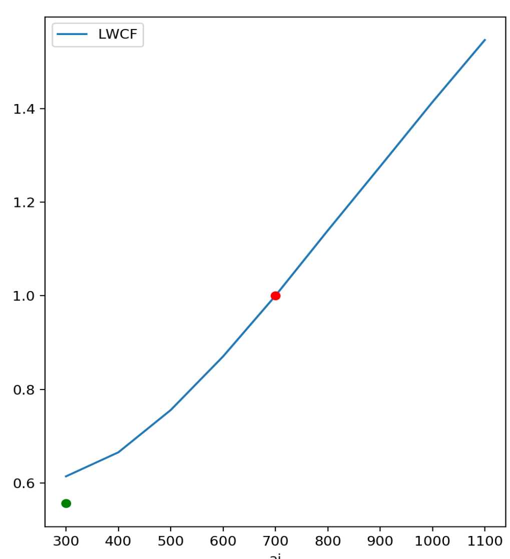
Concerning the auto-tuning itself, the parameter calibration of this work is an unconstrained optimization. It focuses on how to improve the tuning metrics, neglecting the other constraints. In the future work, we plan to use the constrained optimization algorithms, such as the sequential weight increasing factor technique (SWIFT) with the downhill-simplex method, which can conduct the auto tuning directed by the selected metrics, but constrained by other interested quantities in the coupled system.

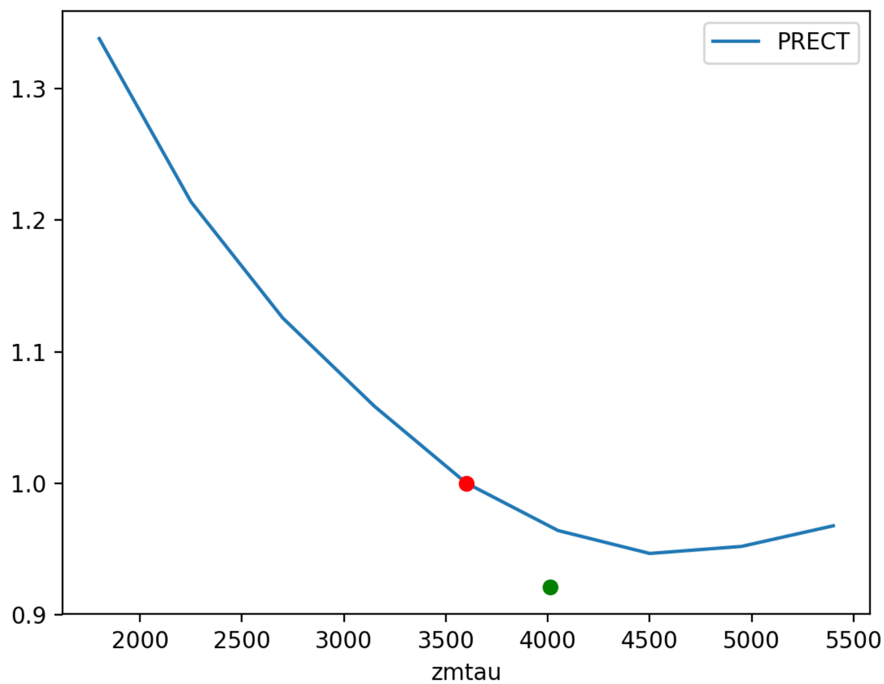
Our tuning strategy does not aim for tuning coupled system, so we do not intend to add this discussion to this manuscript.

4. It would be good to do some of the initialized and maybe AMIP simulations with the individual optimized, to confirm the discussion of the reasons for field improvements.

Reply: Thank you for the suggestion. We try to answer this question by discussing the individual parameter optimization as follows. The two following figures show the individual parameter sampling for LWCF and PRECT in the short-term hindcasts simulation. The y-axis is the relative mean square error, which means that the MSE of the tuned model is normalized by the MSE of the default model, defined in Eq. 1-3 in the manuscript. The red dots correspond to the default parameter values. The green dots are from the optimal obtained in this work.

In the first figure, the relative MSE of LWCF always increase with the increasing sedimentation velocity of ice crystals (ai). The default parameter locates near the middle of the likely range. The smaller ai will lead to a smaller MSE. In the second figure, the relative MSE of PRECT firstly reduces and then nearly stabilize when increasing convective adjustment time scale (zmtau). The bigger zmtau will lead to a smaller MSE. These results confirm the consistency of the automatic multi-parameter tuning in optimizing the parameter towards minimizing the error metric.





5. Incompatible observations versus structural errors: Is it clear to what extend the role of each of these factors maybe. e.g., if you optimized to ERBE or TRMM instead what would be the outcome.

Reply: The following two tables show the biases with respect to different observations for precipitation in terms of RMSE and long-wave cloud forcing in terms of percentage bias, defined as:

$$bias\|\%\| = |\frac{\overline{EXP}(\overline{CNTL}) - \overline{OBS}}{\overline{OBS}}| \times 100\%$$

They show the consistency in reducing RMSE or mean biases with the optimal tuning.

RMSE	Default	Tuned
Precipitation (ERAI)	0.97	0.95
Precipitation (GPCP)	1.20	1.16
Precipitation (TRMM)	1.59	1.51

Bias ||%||	Default	Tuned
LW Cloud Forcing (ISCCP)	17.607	8.643
LW Cloud Forcing (CERES-EBAF)	14.528	5.223

6. Ultimately, does it give greater insight into model tuning that we already have? i.e., if you made a list of parameter shifts required to optimize the model, would it be a surprise to someone who does traditional tuning, and could they give you a reason as to why they would not select these parameter values.

Reply: We think it may not be a surprise to someone who does traditional tuning, but it would be after the fact. It would be hard to imagine with hand tuning to pick the values simultaneously for multiple parameters as quantitatively as derived from auto-tuning.

The traditional tuning bases on the trial-and-error method. Someone who does the traditional tuning usually 1) changes the values of single parameter, 2) analyzes the corresponding response in terms of some fields and 3) determine the tuning direction. This method is subjective and hard to tune multiple parameters simultaneously. The tuning process is rough. The automatic tuning technology use the mathematical optimization algorithm. During each optimal iteration, the algorithm can compute and update the optimal multi-parameter values automatically. This algorithm has been proved that they can find the local minimum or the global minimum solution. These papers, Zhang et al. (2015), Yang et al. (2013) showed the automatic tuning method could further improve the simulation based on the well manual tuned model. The figures in Reply 4 illustrate the multi-parameter automatic tuning tends to be better than the single parameter hand tuning. The more optimized results can help us understand the interactive effect of multi-parameter, discover the systemic and structural errors by exploring the parameter calibration ultimate performance. While there are many benefits of automatic tuning, limitations of this method are also clearly seen. First, the automatic tuning or the traditional tuning are not an alternative to improving the structural errors of models. Second, field improvement by tuning parameters may be caused by compensation of errors.

We have added a sentence to highlight the meaningfulness of the multi-parameter automatic tuning in the manuscript in Section 5.

** Other Comments **

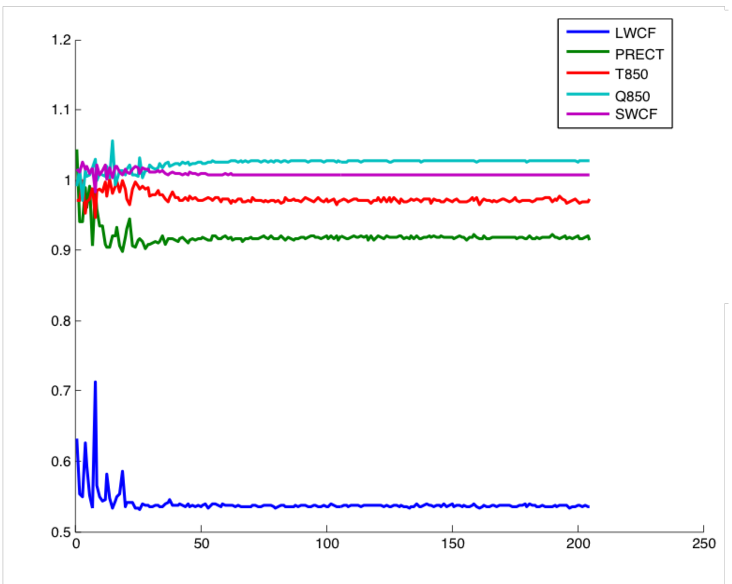
7. Introduction 2. Model and Experiments "The philosophy behind the hindcasts is to keep the model dynamics as close to observation as possible while test how the model simulates the physical variables" -I'm not sure what is meant by this? That you initialize both the dynamics and physics and yet the physics have longer timescales. Do you that this is indeed the case, by looking at RMSE/ACC for dynamical versus physical fields? So the only way you're 'keeping' the dynamics close is by initializing them close to obs. and doing the short runs right? -Could

dynamical fields be a part of the optimization variable set (e.g., U850) to see if they are non-fast? 3-day biases versus -> climate model biases may be different.

Reply: We revised the text to describe the purpose of using short-term hindcasts more clearly in Section 2.

8. Tuning metrics Equation 1 and 2 punishes bad performing variables disproportionately, such that improvements can dominate all other improvements/degradations. This is true for Taylor Diagram scores. Fig 3, would be great if you plotted the individual optimized fields to see how they evolve and iterate to a minimum (probably SWCF is little change and LWCF dominates the optimization).

Reply: The normalized MSE for individual optimized fields are shown in the below figure for reference. We decided to keep the original Figure 3 as the combined metrics can better represent the overall performance of the optimal simulation. Nevertheless, the optimization based on the combined metric cannot guarantee that all the fields get same improvement; as you indicated, it is dominated by the LWCF. In the future work, we plan to use the multi-objective optimization algorithm, which can hopefully solve this problem.



9. 4.2 Figure 4/5: Although LWCF is improved the regional distributions are not that different in contrast to the control. It's therefore somewhat strange that this is not that different, even though LWCF was a target metric. But high-cloud is much improved even though it wasn't a target metric.

Reply: The regional improvement in LWCF is consistent with the improvements in high clouds. It is not as noticeable largely because of the color scale. The difference in LWCF can be more easily seen in the zonal mean LWCF in Figure 7a. Although the high cloud wasn't a target metric, in reality it dominates the change in LWCF. So it is not a surprise to see the changes in high clouds that are consistent with the changes in LWCF.

10. -In figure 7, it is a challenge to really say that the optimization seems not to be operating on anything but LWCF. What happens if you take LWCF out of the optimization fields? Does it look

very similar to the control or does it find a more effective minimization for the remaining variables. -Also how significant are these control/exp changes in AMIP given the short nature of the runs. -If the models surface is warmer does this also imply a surface (and maybe TOA) positive energy imbalance -I am not sure the structural argument is a given. If it is then, it would imply that this process is not truly useful until we go away and fix these types of formulation problem. Incompatible observations may also be a contributing factor.
Reply: Thank you for the helpful comments. It would indeed be interesting to see how the process would go if taking LWCF out. We would prefer to consider this exploration in a separate work, including using multi-objective optimization with LWCF metrics on a separate objective.

We are not concerned about the length of the AMIP simulations because it is driven by climatology mean SST and sea ice coverage, and the results are not prone to interannual variability of underlying surface conditions. Similar length of simulations is commonly used in the literature for various numerical experiments.

11. Conclusions -If model parameter sections were automated and not restricted to ranges, isn't there a danger that the resulting climate could be improved in certain ways, but at the large expense of non-optimized parameters. For example a climate that may have zero convective precip.
Reply: Yes, it is dangerous if the parameters are not restricted to ranges. The ranges ensure the parameters abide by implied physical principles. In this study, each parameter is restricted to specified ranges, which are based on previously published works (Covey et al., 2013; Yang et. al., 2013; Qian et. al., 2015; Zhang et. al., 2013). They are listed in Table 1.