Dear Editor and Referees,

We really appreciate the careful reading, detailed comments and constructive suggestions by the reviewers, which greatly improve the clarity of our presentation and help our revision.

Responses to comments from Referee 1:

1. *Page 3798, line 17, "quit" should be "quite".*

   Reply: Corrected

2. *Page 3801, the sentence in line 16 is a repeat of the 1st sentence on the same page. It should be removed.*

   Reply: Corrected

3. *Page 3802, line 25: This long sentence is not clear. It should be rewritten.*

   Reply: We are sorry about that. This section has been reorganized for better clarification. The corresponding sentence is changed to "Two extra steps are included before the original downhill simplex method to overcome its limited effectiveness on model performance improvement. The "Downhill_2_steps" method includes an initial value pre-processing step before the downhill simplex method. And the "Downhill_3_steps" method further introduces another step to eliminate insensitive parameters for tuning by sensitivity analysis."

4. *Line 3804, 2nd paragraph: The physical explanations should be improved. If the model used the stratiform fractional cloud condensation scheme of CAM3 or CAM4 (Zhang, et al. 2003), reducing the "rhminh" threshold will not only increase the cloud amount, but also increase the stratiform condensation rate and decrease the atmospheric humidity. Likewise, increasing the "rhminl" will do the opposite. This is why you see clear opposite changes of RH and CLOUD in the lower troposphere and upper troposphere in Figure 6.*

Reply: We thank the reviewer for pointing out this important linkage and the illuminating explanation for Fig. 6. Yes, the stratiform fractional cloud condensation scheme of CAM4 (Zhang, et al.2003) was used in GAMIL2. Accordingly, we have included the following to the model description section on page 7, line 5. At the same time, we rewrite the 1st paragraph on page 15 of the revised manuscript.

"Compared to the previous version, GAMIL2 has modifications in cloud-related processes (Li et al. 2013), such as the deep convection parameterization (Zhang and Mu, 2005), the convective cloud fraction (Xu and Krueger, 1991), the cloud microphysics (Morrison and Gettelman, 2008), and the stratiform fractional cloud condensation scheme (Zhang et al. 2003)."

"With reduced RH threshold for high cloud (from 0.78 in CNTL to 0.63 in EXP, Table 1), the stratiform condensation rate increases and the atmospheric humidity decreases (Zhang et al. 2003). In addition, with increased auto-conversion coefficient in the deep convection, less condensate is detrained to the environment. As a result, mid- and upper-troposphere is overall drier, especially over the tropics where deep convection dominates the vertical moisture transport (Fig. 6c). Although the mid- and upper-troposphere become drier over the tropics, reduced RH threshold for high cloud makes clouds easier to be present. Consequently, middle and high clouds increase over the globe, especially over the mid- and high-latitudes with the largest increase up to 4–5 % (Fig. 6f). In the tropics, due to the drier tendency induced by the reduced detrainment, high cloud increase is relatively small (2–3%) compared to the mid- and high-latitudes. On the contrary, low cloud below 800 hPa decreases by 1–2% over the mid- and high-latitudes with slightly decreased RH (Fig.6) because of the negligible change of RH threshold for low cloud (Table 1). Overall, the combined effects of all relevant parameterizations lead to the changes of atmospheric humidity and cloud fraction."

Responses to comments from Referee 2:

General comments: This study proposed a "three-step" parameter optimization procedure which can help tuning important parameters in general climate models with reduced computation load. This "three-step" procedure is an extension of downhill simplex method with a parameter sensitivity process to eliminate insensitive parameters and an initial value selection process to help improving optimization converging quality. Results show that by finding an optimal set of parameter values, the method is able to improve the climate simulation compared with default parameter values. At the same time, the computation time required is reduced compared with traditional methods. However, there are great deficiencies in illustrating the methodology. Both the core procedure downhill simplex method and the extended parameter sensitivity process and initial value process are not clearly presented, making it very difficult for readers to follow and learn. Also, there is not enough meaningful comparisons between the results of new method and those of traditional methods for readers to judge whether it is a progressive method. A future version of this manuscript may potentially be acceptable. But that apparently requires a lot more work.

Reply: We agree with the reviewer that the description of the method is not clear. We have substantially revised this part for better clarification and presentation. More details can be found in the revised manuscript and the point to point responses to the reviewer. First of all, the use of "local vs global" in several places has induced confusions. There are two groups of usage of "local vs global" in the manuscript. The first one refers to the parameter sensitivity, in which "local" means the model's sensitivity to a single parameter and "global" means the model's sensitivity to all the parameters in consideration. The second one refers to the optimization methods, in which "local" means the method searching for a local optimum solution and "global" means the method aiming for the global optimum solution. We have thought to change the first use of "local vs global" to some other nomenclatures, such as "single parameter sensitivity or combined parameter sensitivity" for better clarification. However, it is a common practice to use "local vs global" in statistics and sensitivity analysis and so we keep them

in the manuscript. Nevertheless, we have paid special attention to the presentation for the clarification in the revised manuscript.

Second, some confusion comes from the usage of special words from mathematics and computer science. These words include “trajectory”, “distance”, “simplex”, “dimension”, among others. We have tried our best to give a brief explanation or description of these words in the text. It is hoped that it will help readers for better understanding.

Regarding the comparison of the new method and traditional methods, we have not illustrated the progress clearly, especially about the explanation of Table 3 and 4. Two performance criteria are used to evaluate the effectiveness and efficiency of the optimization algorithms in this study. Selection of optimization algorithms for parameter calibration of climate system models is a balance between model improvement (effectiveness) and computational cost (efficiency). In this study, model improvement is measured by an index defined in Eq. (3). The lower of this value is, the better model tuning is. Computational cost is measured by "core-hours", standing for the computational efficiency. It is computed by (N_{step}) * (N_{size}) * (the number of process of a single model run) * (hours used for a single 5-year model run). N_{step} is the total numbers of iterations of optimization algorithms for convergence. N_{size} is the number of model runs during each iteration, and it is 1 for the downhill simplex method.

Effectiveness and efficiency of the three traditional algorithms are compared in Table 3. “Downhill_1_step” represents the original downhill simplex method, which is one of the most widely-used local optimization algorithms and has been successfully used in Speedy model (Severijns and Hazeleger, 2005). PSO and DE are the most widely-used global optimization algorithms and easy-to-use. Although “Downhill_1_step” achieves slightly worse improvement compared to the two global optimization methods (Table 3), its computation cost is much less (only 20% and 28% of DE and PSO respectively). The most important contribution of the study is adding two extra steps to the original downhill simplex method. We are able to achieve better improvement with less computational cost than the two global methods (Table 4). The “Downhill_2_steps” method includes the initial value pre-processing before the downhill simplex method.

And the "Downhill_3_steps" method further introduces an extra step to determine parameters for tuning by sensitivity analysis. Table 3 and 4 show that the proposed "Downhill_3_steps" is able to overcome the inherent ineffectiveness of the original downhill simplex method with much lower computational cost than global methods. We have clarified and emphasized this in the revised manuscript.

In addition, a comparison of the CNTL and EXP is used to illustrate how the tuning of these parameters improves the model results in terms of various atmospheric fields. This helps the readers for a better understanding of the physical reasons behind the automatic tuning process.

Once again, we thank the reviewer for his/her effort and time to help us improve the manuscript. A lot of works have been done to develop the methodology and we hope it would be a useful tool for the model development community.

Point to point responses:

1. *Page 3792, Line 9: "parameter sensitivity" should be more specified, such as the model's sensitivity to the parameters. "optimum initial value" should be specified for the parameter estimation process.*

   Reply: This sentence has been rewritten as "Different from the traditional optimization methods, two extra steps, one determining the model's sensitivity to the parameters and the other choosing the optimum initial value for those sensitive parameters, are introduced before the downhill simplex method. This new method reduces the number of parameters to be tuned and accelerates the convergence of the downhill simplex method. "

2. *Page 3792, (3794?) Line 10: What does the "step" refer to? Parameter optimization cycles? Model integration steps? Or method cycles?*

   Reply: The "step" here refers to the optimization cycles involved within the optimization algorithm.

3. *Page 3794, Line 3: "high" should be "high-dimensional".*

Reply: Corrected.

4. *Page 3794, Line 19-20: ENKF and PF have the difficulty in looking for the representative samples: This problem needs to be explained more clearly and needs to be extended a little, and references should be introduced.*

   Reply: The sentence "ENKF and PF have the difficulty in looking for the representative samples" has been rewritten as "The EnKF and PF use an ensemble of model simulations to estimate the background error covariance, which approximate the traditional Kalman filter with a recurrence process (Evensen 2003, Arulampalam 2002). The accuracy of the error covariance relies on samples. In general, the larger the ensemble size, the more accurate the estimates are. The limitation of ensemble size for practice use and imperfect models make it difficult to select representative samples (Poterjoy 2014). "

   References:

   Evensen G. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean dynamics, 2003, 53(4): 343-367.

   Arulampalam M S, Maskell S, Gordon N, et al. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. Signal Processing, IEEE Transactions on, 2002, 50(2): 174-188.

   Poterjoy J, Zhang F, Weng Y. The effects of sampling errors on the EnKF assimilation of inner-core hurricane observations. Monthly Weather Review, 2014, 142(4): 1609-1630.

5. *Page 3794, Line 25-26: "The above mentioned methods generally require long iterations for convergence." This is not necessary. It also depends on observation amounts and estimation frequencies.*

   Reply: This sentence has been deleted since it does not fit in the paragraph.

6. *Page 3795, Line 8-10: "Finally, the downhill simplex algorithm is used to solve the optimization problem because of its low computational cost and fast convergence for low dimension space." This dimension space corresponds to parameter space? Also, you have said that the parameter space in climate models are usually high-dimensional. Does it mean that this method is not suitable for climate model tuning?*

   Reply: Yes, here dimensional space means parametric space. The default downhill simplex is not good at high-dimensional parametric space. This is why we propose a "three-step" method to reduce the number of parameters (i.e., dimension of space) by sensitivity analysis. The results of Table 4 show that the "three-step" method is able to achieve the best tuning results compared to other tuning algorithms with relatively low computational cost.

7. *Page 3795, Line 12. "This is result already." What does it mean here?*

   Reply: This is a typo. We have deleted this sentence.

8. *Page 3797, Line 13-14: "Previous studies have shown 5 years of this type of simulation is enough to capture some basic model characteristics." What are these basic model characteristics? Should be extended and necessary references should be included.*

   Reply: This sentence is rewritten as "Previous studies have shown 5 years of this type of simulation is enough to capture the basic characteristics of simulated mean climate states (Golaz et al. 2011, Lin et al. 2013)".

   References:

Golaz J C, Salzmann M, Donner L J, et al. Sensitivity of the aerosol indirect effect to subgrid variability in the cloud parameterization of the GFDL atmosphere general circulation model AM3[J]. Journal of Climate, 2011, 24(13): 3145-3160.

Lin Y, Zhao M, Ming Y, et al. Precipitation partitioning, tropical clouds, and intraseasonal variability in GFDL AM2[J]. Journal of Climate, 2013, 26(15): 5453-5466.

9. *Page 3797, Line 17: "reference metrics". what is this metric like? It is a metric containing those climate variables? How is it formed?*

   Reply: Thanks for pointing out the confusing presentation. The model performance during parameter tuning process depends on the metrics used for the evaluation. In this study, a basic metrics including the mean states of wind, humidity, geopotential height field, and various radiative fields, is used for illustration. Note that the metrics can be easily expanded. Page 3797, Line 17: "reference metrics" and Page 3797, Line 24: "evaluation metrics" are the same. And we changed the evaluation metrics to reference metrics in line 24. Accordingly, "A comprehensive metrics,…" in line 26 is changed to "the reference metrics, ". The metrics is described at Page 8, Eq. (1) ~ (3). It is an improvement index to evaluate the tuning results, which weight each variable equally and compute the average normalized RMSE. The index indicates an overall improvement of the performance of the tuned simulation relative to the control simulation according to a number of model outputs (Table 2). If the index is less than 1, it means the tuned simulation gets better performance than the control run. The smaller this value, the better improvement is.

10. *Page 3797, Line 24: "evaluation metrics". What is the difference between the reference metric and evaluation metric? What is this evaluation metric like again?*

    Reply: Please see the 9th reply.

11. *Page 3797, Line 26: "metrics". So this metric is the evaluation metric?*

    Reply: Please see the 9th reply.

12. *Page 3798, Line 2: "control simulation". What is this control simulation here? With default parameter values? Please specify.*

Reply: Yes, the control simulation refers to the simulation using default parameter values. The sentence is changed to "we normalize the RMSE of each simulation output by that of the control simulation using default parameter values."

13. *Page 3798, Line 10: "w is the weight due to the different grid area". What is w like? Is it the same weight?*

Reply: This is because the model output is on regular latitude longitude grids, which have varying grid areas. The grid weight (w) is computed as cos(the latitude of each grid) to consider the area change of different grid cells. The sentence is changed to "w is the weight due to the different grid area on regular latitude longitude grids on the sphere."

14. *Page 3798, Line 13: "Global and local optimization method." This section is supposed to tell the methodology of global and local optimization method. But the authors only listed typical examples and names of each kind without explaining the methodology. The whole section is rather too simplified that it is difficult to understand.*

Reply: A nice point. The title of section 4.1 is changed to "Parameters tuning with global and local optimization methods". We revise the first paragraph and summarize the main differences between the global and local optimization methods. The first paragraph has been rewritten as:

"Parameter tuning for a climate system model is to solve a global optimization problem in theory. As the well-known global optimization algorithms, traditional evolutionary algorithms, such as genetic algorithm (Goldberg et al., 1989), differential evolutionary (DE) (Storn and Price, 1995), and particle swarm optimization (PSO) (Kennedy, 2010), can approach the global optimal solution but generally require high computational cost. This is because these algorithms are designed following biological evolution of survival of the fittest. In contrast, the local algorithms utilize the greedy strategy, and thus may stick at a locally optimal solution after convergence. The advantage of local algorithms is the low computational cost due to relatively less

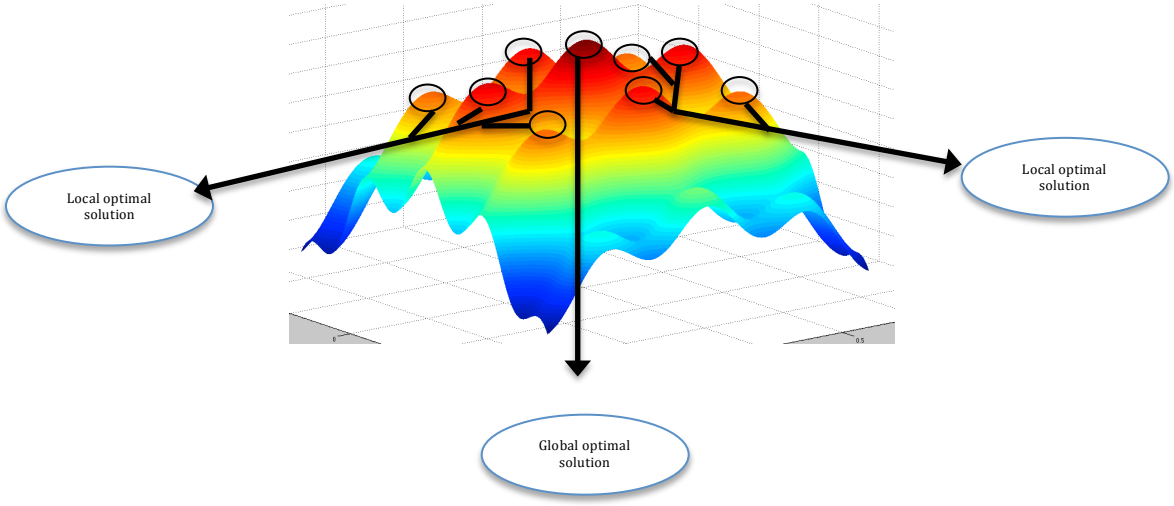
samples required. In this sense, the local optimization algorithms are the viable options considering their significantly reduced computational cost."

15. *Page 3799, Line 14: "local downhill simplex method". So the local downhill simplex method gives an optimal parameter value sets locally? Say, each region has an optimal set? And these parameter sets are assigned into next model integration cycle locally? Could you add some explanation about the specific methodology of downhill simplex method? And what is the difference between a local optimization and global optimization? If local, then local to where? spatially or in other space? If this "local" refers to spatial local concept, then why in Table 3 the optimization is one value set? Is it because it is local optimization to some specific region? Also without a clear explanation of the methodology of "local" and "global" method, there is no way for readers to understand the results and why global optimization gives better tuning results. And how do you judge "better" results. There is no direct comparison with a certain reference criterion here.*

Reply: Thanks for pointing out the confusion. "local" and "global" here does not refer to the spatial locations of the GCM, but the local or global optimum solution algorithms can approach as emphasized in the summary section of this response. We add the following sentences for a better clarification of "local vs. global". The "local" and "global" refer to that the optimization algorithms can achieve the "local" or "global" convergence performance. For example in the following figure, the horizontal coordinate stands for the two-dimension parametric space, and the vertical coordinate presents the metrics. If the optimal target is the maximum metrics, there are many local optimal solutions and one global optimal solution.

The downhill simplex is a local method and can be trapped within a local solution space. Instead, the global methods, such as genetic algorithm, DE and PSO, are able to find the global optimal solution. However, the local method is faster and requires much less model runs than the global method. In this study, we try to overcome the problems of local method (the downhill simplex) by introducing two extra steps: parameter sensitivity analysis and initial values selection. With relatively low

computational cost, this new method can reduce the number of parameters to be tuned, determine the local parametric space for final solution, and accelerate the convergence of the downhill simplex method.



The revised version adds the following text to explain the estimation criteria in the third paragraph of section 4.1 before the sentence "According to tuning GAMIL2…":

"Two performance criteria are used to evaluate the effectiveness and efficiency of the optimization algorithms in this study. Selection of optimization algorithms for parameter calibration of climate system models is a balance between model improvement (effectiveness) and computational cost (efficiency). In this study, model improvement is measured by an index defined in Eq. (3). The lower of this value is, the better model tuning is. Computational cost is measured by "core-hours", standing for the computational efficiency. It is computed by (N_{step}) * (N_{size}) * (the number of processes of a single model run) * (hours used for a single 5-year model run). N_{step} is the total numbers of iterations of optimization algorithms for convergence. N_{size} is the number of model runs during each iteration, and it is 1 for the downhill simplex method. In GAMIL2 case, each model run takes 6 hours using 30 processes. "

16. *Page 3800, Line 6: "The Morris method". At least a simple explanation about this "global" method should be provided. So it is a method of perturbing all parameters? After reading this paragraph, I still didn't get how the sensitivity experiment is carried out.*

Reply: Thanks for the nice point. The second and third paragraphs in Section 4.2 has been rewritten as:

"Parameter sensitivity analysis can be divided into local and global methods (Gan et al., 2014). The local method determines the sensitivity of a single parameter by perturbing one parameter with all other parameters fixed. Consequently, it does not consider the combined sensitivity of multiple parameters. On the other hand, the global method perturbs all the parameters to explore the sensitivity of the whole parametric space. In this study, the Morris (Morris, 1991; Campolongo et al., 2007), a global method, is used to screen out the sensitive parameters. Another global method (Sobol 2001), is used to validate the results of the Morris method.

The Morris method, based on the MOAT sampling strategy, reduces the number of samples required by other global sensitivity methods (Li et al., 2013). Note that a sample is a set of all parameters, not just one parameter. The method is described briefly here, and more details can be found in Morris (1991). Assume we have k parameters relative to a random sample $S_1 = \{ x_1, x_2, \ldots, x_k \}$, another sample $S_2 = \{ x_1, x_2, \ldots, x_i + \Delta_i, \ldots, x_k \}$ can be constructed by perturbing the i_{th} parameter by Δ_i, where Δ_i is a perturbation of this parameter. The elementary effect of the i_{th} parameter x_i is defined as:

$$d_i = \frac{f(S_2) - f(S_1)}{\Delta_i} \qquad (4)$$

where f stands for the improvement index as defined in Eq. (3). A third sample $S_3 = \{ x_1, x_2, \ldots, x_i + \Delta_i, \ldots, x_j + \Delta_j, x_k \}$ can be generated by perturbing another parameter. In so doing k times, we will get k+1 samples $\{S_1, S_2, \ldots, S_{k+1}\}$, and k elementary effects $\{d_1, d_2, \ldots, d_k\}$ after perturbing all the parameters. The vector of $\{S_1, S_2, \ldots, S_{k+1}\}$ is called a trajectory. This procedure is repeated for r iterations and finally we get r trajectories. The starting point of any trajectory is selected randomly as well as the ordering of the parameters to perturb and the Δ for each perturbation in one trajectory. In practice, a number of 10 to 50 trajectories are enough to determine

the feasible sensitivity of parameters (Gan et al., 2014; Morris, 1991). In this study, we have a total of 7 parameters and 80 simulations are conducted.

We define D = $\{d_i(t)\}$, where t is the t_{th} trajectory, and i is the i_{th} elementary effect of the parameter x_i. μ_i, the mean of $|d_i|$, and σ_i, the standard deviation of d_i, are used to measure the parameter sensitivity, defined as:

$$\mu_i = \sum_{t=1}^{r} \frac{|d_i(t)|}{r} \quad (5)$$

$$\sigma_i = \sum_{t=1}^{r} \sqrt{(d_i(t) - \mu_i)^2/r} \quad (6)$$

μ_i estimates the effect of x_i on the model improvement, while σ_i assesses the interactive effect of x_i with other parameters. Those parameters with large μ_i and σ_i are the sensitive parameters. The Morris method results are shown in Fig. 2."

17. *Page 3800, Line 11-12: "n is the number of calibration parameters". Is n different from N? If so, a consistent denotation should be used. What does it mean by trajectories? Sample simulations?*

Reply: This is a typo. "n" is the "N". Also, we can find the description of trajectory in the 16th reply.

18. *Page 3800, Line 14: "step size". What do you mean by step size here? Number of integration steps?*

Reply: This is a typo. It is "perturbation". For more details, please see the 16th reply.

19. *Page 3800, Line 15: "The starting point of a trajectory is selected randomly". What do you mean by trajectory? How do you provide initial condition? How is the parameter initial values chosen? Randomly? If the parameter initial values are chosen randomly, it is not convincing that the randomly given values would give accurate estimation of parameter sensitivity. And for sure it would take a very long*

time for parameter optimization to converge. And it is highly likely that the parameter would converge to a total wrong value.

Reply: Please see the 16th reply for the description about the trajectory. As in the 16th reply, the i_{th} elementary effect of the parameter x_i only perturbs the x_i with other parameters fixed in a trajectory. If the starting point of any trajectory is selected randomly, it can ensure that all parameters can be perturbed in different trajectories. Therefore, we get the global sensitive results with the Morris method.

The sensitivity analysis is used to determine the sensitive parameters. After this step, we begin to tune the parameters by optimization algorithms. The wrong results of parameter screening may lead to low quality solution in the optimization step.

20. *Page 3800, Line 22: what is y? How to choose the integration time? Because after changing a parameter, the model would shortly respond in a linear manner and later exhibit nonlinear response? How to choose the integration time to compare y? Besides, how do you choose the parameter step size? Based on what?*

Reply: In this paper, y is the improvement index as defined in Eq. (3). The atmosphere simulations are conducted for 5 years. We describe it in Page 8. Previous studies have shown 5 years of simulation is enough to capture the response of climate mean states. The "parameter step size" means the parameter perturbation, which is an integer multiple of the discrete perturbation scale mentioned in the 16th reply. Please see the revised text for more details.

21. *Page 3800, Line 25: I didn't see any sensitivity results in Fig.1. It should be Fig. 2.*

Reply: Corrected. It is changed to Fig.2.

22. *Page 3800, Line 25-27. The model's sensitivity to the parameters is somehow dependent on the perturbation magnitude. In terms of response time, model can be very sensitive to some parameters that the quickly displayed spread. However, to some parameters, the model's response is rather slow. In terms of sensitivity magnitude, the model could respond to the parameter, however, the magnitude of the*

spread could be small. In your study, it seems that only the magnitude is included as a criterion of sensitivity. And the parameter perturbation, step size, is not well explained here.

Reply: Thanks for pointing out this. If we understand correctly, the reviewer suggests we should take both the model response magnitude and the response time into account here. It depends on the design of metrics. As an example, in this study we use the mean states of different model outputs as the metrics, an improvement index to evaluate the tuning performance, and the parameters with larger response in 5-year simulations will be chosen for during tuning. If we want to take the response time into account, we need to design a new metrics first.

Moreover, since all the parameters tested in this study are related to cloud and convection, which are generally called the fast physics and their impacts on model climate states will manifest quickly. In addition, atmospheric-only simulations do not involve ocean and other slow components of the climate system. A few years of such types of simulation are long enough to capture the overall climate states as measured by the defined metrics.

The detail explanations for "parameter perturbation" and "step size" can be found in the 16th reply.

23. *Page 3801, Line 1-10: This paragraph seems to be an old version of the next paragraph.*

Reply: Corrected.

24. *Page 3801, Line 24-25: Why is that? As I understand from your previous description, "local" here means the model's response to one single parameter. And this does not necessarily lead to a dependence on the initial value.*

Reply: "local" here refers to the local optimization algorithm, not the model's response to one single parameter. For a local optimization algorithm, it only searches for a local optimal solution and its convergence strongly depends on the initial value

of the parameter. Actually, some bad initial values of parameters may lead to non-convergence for the local method. We have clarified this in the revision version.

25. *Page 3802, Line 4: what do you mean by a longer distance? What is the distance? Compared to what it is longer?*

Reply: We are sorry about the confusion here. The first and second paragraphs have been rewritten as:

"The downhill simplex method is a local optimization algorithm and its convergence performance strongly depends on the quality of the initial values. We need to find the parameters with the smaller metrics around the final solution. Moreover, we have to finish the searching as fast as possible with minimal overhead. For these two objectives, a hierarchical sampling strategy based on the single parameter perturbation (SPP) sample method is used. The SPP is similar to local sensitivity methods, in which only one parameter is perturbed at one time with other parameters fixed. The perturbation samples are uniformly distributed across parametric space. First, the improvement index as defined in Eq. (3) of each parameter sample is computed. The distance is defined as the difference between the improvement indexes using two adjacent samples, i.e., the model response measured by certain percentage change of one parameter. We call this step the first level sampling. The specific perturbation size for one parameter can be set based on user experience. In our implementation, user needs to set the number of samples. For the first level sampling, we can use a larger perturbation size to reduce computational cost. If the distance between two adjacent samples is greater than a predefined threshold, more SPP samples between the previous two adjacent samples are conducted. And this is called the second level sampling. Finally, k+1 samples with the best improvement index value are chosen as the candidate initial values for the optimization method. With this hierarchical sampling strategy, we can determine the local parametric space for final solution and can accelerate the convergence of the following downhill simplex method. This procedure is described in Algorithm 1. It is easy to implement and has lower overhead compared to other complex adaptive sampling methods.

Algorithm 1 Preprocessing the initial values of Downhill Simplex Algorithm.

```
//Single parameter perturbation sample(SPP)
N=number_of_parameters
sampling_sets={}
for each parameter P_i of N parameters do
    sampling_sets+=SPP(P_i_range, number_of_samples)
    //refine sample in the sensitivity range if needed
    if metrics of the the adjacent same parameter sampling points >= sensitivity_threhold then
        sampling_sets+=SPP(P_i_adjacent_parameter_range, refine_ number_of_factors)
    end if
end for

//Initial vertexes with parameters of the N+1 minimum metrics
for each initial V_i of N+1 vertexes do
    //get the parameters of the ith minimum metrics
    candidate_init_sets += min(i, sampling_sets)
end for

//make sure the initial simplex geometry is well-conditioned
while one parameter k have the same values in the N+1 sets do
    j = 1
    //remove the parameter set with the worst metrics from candidate_init_sets
    remove_parameter_set(the parameter set with worse metrics, candidate_init_sets)
    //get the parameters of the N+1+jth minimum metrics
    candidate_init_sets += min(N+1+j, sampling_sets)
    j+=1
end while
```

At the same time, inappropriate initial values may lead to ill-conditioned simplex geometry, which can be found in model tuning process. One issue we meet is that some vertexes in the downhill simplex optimization may have the same values for one or more parameters. As a result, these parameters remain invariant during the optimization and this may degrade the quality of final solution as well as the convergence speed. A simplex checking is conducted to keep as many different values of parameters as possible during the process of looking for initial values. Well-conditioned simplex geometry will increase the parameter freedom for optimization. In our implementation (Algorithm 1), the vertex leading to the ill-conditioned simplex is replaced by another parameter sample which gives another minimum improvement index value."

26. *Page 3802, Line 5-6: "a smaller distance". I don't understand the distance here? Is it represented by any denotations in the Equations listed before?*

Reply: Please see the 25th reply.

27. Page 3802, section 4.3: After reading the whole section, I still cannot get how to get the inital value.

Reply: Please see the 25th reply.

28. Page 3802, Line 22-23: "In Table 3, PSO gets the best solution." How do you get this conclusion? Can you provide any reference parameter value or error information so that we can tell which estimation is the best?

Reply: We are sorry about the confusion here. The conclusion comes from the revised third paragraph of section 4.1.

"Two performance criteria are used to evaluate the effectiveness and efficiency of the optimization algorithms in this study. Selection of optimization algorithms for parameter calibration of climate system models is a balance between model improvement (effectiveness) and computational cost (efficiency). In this study, model improvement is measured by an index defined in Eq. (3). The lower of this value is, the better model tuning is. Computational cost is measured by "core-hours", standing for the computational efficiency. It is computed by (N_{step}) * (N_{size}) * (the number of processes of a single model run) * (hours used for a single 5-year model run). N_{step} is the total numbers of iterations of optimization algorithms for convergence. N_{size} is the number of model runs during each iteration, and it is 1 for the downhill simplex method. In GAMIL2 case, each model run takes 6 hours using 30 processes. "

PSO has the lowest "final optimal model metrics", meaning that it gets the best effective solution compared with other traditional methods in Table 3. This sentence is changed to "In Table 3, PSO gets the best effective solution."

29. Page 3803, Line 2-3: I still didn't get how you judge whether this estimation is good or bad.

Reply: As in the 15th and 28th responses, the "final optimal model metrics" is used to estimate the model improvement (effectiveness), and the "core-hours" for

computational cost (efficiency). Selection of optimization algorithms for parameter calibration of climate system models is a balance between effectiveness and computational cost efficiency. Effectiveness and efficiency of the three traditional algorithms are compared in Table 3. "Downhill_1_step" represents the original downhill simplex method, which is one of the most widely-used local optimization algorithms and has been successfully used in Speedy model (Severijns and Hazeleger, 2005). PSO and DE are the most widely-used global optimization algorithms and easy-to-use. Although "Downhill_1_step" achieves slightly worse improvement compared to the two global optimization methods (Table 3), its computation cost is much less (only 20% and 28% of DE and PSO respectively). The most important contribution of the study is that by adding two extra steps to the original downhill simplex method, we are able to achieve better improvement with less computational cost than the two global methods (Table 4). The "Downhill_2_steps" method includes the initial value pre-processing before the downhill simplex method. And the "Downhill_3_steps" method further introduces an extra step to determine parameters for tuning by sensitivity analysis. Table 3 and 4 show that the proposed "Downhill_3_steps" is able to overcome the inherent ineffectiveness of the original downhill simplex method with much lower computational cost than global methods. Therefore, our proposed method has a good trade-off between accuracy and computational cost.

30. *Page 3803, Line 21: "The change in terms of the RMSE factor". So how to calculate this change in RMSE? What RMSE quantity is shown in FIg.5?*

Reply: The RMSE is the metrics described at Page 8, Eq. (1) ~ (3) in the revised manuscript.

$$(\sigma_{\mathrm{m}}^{F})^2 = \sum_{i=1}^{I} w(i)(x_{\mathrm{m}}^{F}(i) - x_{\mathrm{o}}^{F}(i))^2 \quad (1)$$

$$(\sigma_{r}^{F})^2 = \sum_{i=1}^{I} w(i)(x_{r}^{F}(i) - x_{\mathrm{o}}^{F}(i))^2 \quad (2)$$

$$\chi^2 = \frac{1}{N^F} \sum_{F=1}^{N^F} \left(\frac{\sigma_{\mathrm{m}}^{F}}{\sigma_{r}^{F}}\right)^2 \quad (3)$$

x_m^F is the model outputs, and x_o^F is the corresponding observation or reanalysis data. x_r^F is the model outputs from the control simulation using the default values for the parameters in Table 1. w is the weight due to the different grid area on a regular latitude longitude grids on the sphere. I is the total grid number in model. N^F is the number of the chosen variables.

Eq. (3) thus defines an improvement index. If the index is less than 1, it means the tuned simulation gets better performance than the control run based on the reference metrics (Table 2). The smaller this value, the better improvement is.

31. *Page 3803, Line 26-27: Maybe, but temperature obs is also included as a criterion in parameter optimization. It is possible that the compromising result will degrade the simulation of temperature, but it is still not very convincing...Have you checked the temperature's and other varibales' sensitivity to the parameters? If the sensitivity of temperature is much smaller than those of others, it may help support your arguement...*

Reply: We deleted this sentence since winds and temperatures are also closely influenced by these parameters. Because the improvement is evaluated by the metrics consisted of all 16 variables, it is possible that some variables become worse in EXP than in CNTL.

32. *Page 3805, Line 14-15: There is no standard criterion for the readers to judge whether the estimation is good or bad.*

Reply: Please see the 29th reply.

33. *Page 3805, Line 21-23: References should be included here. However, the surrogate-based optimization method seems to have no relation with this study at all, thus inappropriate to be formed as a comparison.*

Reply: Yes. The surrogate-based optimization method is not addressed in this study and is still under investigated by the authors as well as other scientists. Moreover, the

proposed method in this paper can also work well together with surrogate models for climate system models. We just delete this paragraph in the revised manuscript.

34. *Page 3805, Line 25-27: Since you have said that the surrogate-based method cannot meet the requirement of climate systems, simply stating that future work focus on evaluate surrogate models seems not very relevant with this study, nor as an extension of this study. More justification is needed.*

Reply: Please see the 33th reply.

35. *Figures: To justify that the three-step method is more effective and more efficient, more comparisons between this new method and the traditional method should be provided. Only comparing between EXP and CNTL is not enough.*

Reply: The effectiveness and efficiency of the three-step method and the comparison with other methods are illustrated in Table 3 and 4. As in the 15th and 28th responses, the "final optimal model metrics" is used to estimate the model improvement (effectiveness), and the "core-hours" for computational cost (efficiency). Selection of optimization algorithms for parameter calibration of climate system models is a balance between effectiveness and computational cost efficiency. Effectiveness and efficiency of the three traditional algorithms are compared in Table 3. "Downhill_1_step" represents the original downhill simplex method, which is one of the most widely-used local optimization algorithms and has been successfully used in Speedy model (Severijns and Hazeleger, 2005). PSO and DE are the most widely-used global optimization algorithms and easy-to-use. Although "Downhill_1_step" achieves slightly worse improvement compared to the two global optimization methods (Table 3), its computation cost is much less (only 20% and 28% of DE and PSO respectively). The most important contribution of the study is that by adding two extra steps to the original downhill simplex method, we are able to achieve better improvement with less computational cost than the two global methods (Table 4). The "Downhill_2_steps" method includes the initial value pre-processing before the downhill simplex method. And the "Downhill_3_steps" method further introduces an extra step to determine parameters for tuning by sensitivity analysis.

Table 3 and 4 show that the proposed “Downhill_3_steps” is able to overcome the inherent ineffectiveness of the original downhill simplex method with much lower computational cost than global methods. Therefore, our proposed method has a good trade-off between accuracy and computational cost.

In addition, a comparison of the CNTL and EXP is used to illustrate how the tuning of these parameters improves the model results in terms of various atmospheric fields. This helps the readers for a better understanding of the physical reasons behind the automatic tuning process.

Manuscript prepared for Geosci. Model Dev. Discuss.
with version 2015/04/24 7.83 Copernicus papers of the LaTeX class copernicus.cls.
Date: 27 August 2015

An automatic and effective parameter optimization method for model tuning

T. Zhang[1,2], L. Li[3], Y. Lin[2], W. Xue[1,2], F. Xie[3], H. Xu[1], and X. Huang[1,2]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Center for Earth System Science, Ministry of Education Key Laboratory for Earth System Modeling, Tsinghua University, Beijing 100084, China
[3]State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China

Correspondence to: W. Xue (xuewei@mail.tsinghua.edu.cn)

Abstract

Physical parameterizations in General Circulation Models (GCMs), having various uncertain parameters, greatly impact model performance and model climate sensitivity. Traditional manual and empirical tuning of these parameters is time consuming and ineffective. In this study, a "three-step" methodology is proposed to automatically and effectively obtain the optimum combination of some key parameters in cloud and convective parameterizations according to a comprehensive objective evaluation metrics. Different from the traditional optimization methods, two extra steps, one determining the model's sensitivity to the parameters and the other choosing the optimum initial value for those sensitive parameters, are introduced before the downhill simplex method. This new method reduces the number of parameters to be tuned and accelerates the convergence of the downhill simplex method. Atmospheric GCM simulation results show that the optimum combination of these parameters determined using this method is able to improve the model's overall performance by 9 %. The proposed methodology and software framework can be easily applied to other GCMs to speed up the model development process, especially regarding unavoidable comprehensive parameters tuning during the model development stage.

1 Introduction

Due to their current relatively low model resolutions, General Circulation Models (GCMs) need to parameterize various sub-grid scale processes. Physical parameterizations aim to approximate the overall statistical outcomes of various sub-grid scale physics (Williams, 2005). However, due to the complexities involved in these processes, parameterizations representing sub-grid scale physical processes unavoidably involve some empirical or statistical parameters (Hack et al., 1994), especially within cloud and convective parameterizations. Consequently, these parameterizations introduce uncertainties to climate simulations using GCMs (Warren and Schneider, 1979). In general, these uncertain parameters need

to be calibrated or constrained when new parameterization schemes are developed and integrated into models (Li et al., 2013).

Traditionally, the uncertain parameters are manually tuned by comprehensive comparisons of model simulations with available observations. Such an approach is subjective, labor intensive, and hard to be extended (Hakkarainen et al., 2012; Allen et al., 2000). By contrast, the automatic parameter calibration techniques have progressed quickly because of their efficiency, effectiveness and broader applications (Bardenet et al., 2013; Elkinton et al., 2008; Jakumeit et al., 2005; Chen et al., 1999). In previous studies applying to GCMs, the methods can be categorized into three major types based on probability distribution function (PDF) method, optimization algorithms, and data assimilation techniques.

For the PDF method, the confidence ranges of the optimization parameters are evaluated based on likelihood and Bayesian estimation. Cameron et al. (1999) improves the forecast by the generalized likelihood uncertainty estimation (Beven and Binley, 1992), a method obtaining parameter uncertain ranges of a specific confidence level. The Bayesian Markov Chain Monte Carlo (MCMC) (Gilks, 1995) is widely used to obtain posterior probability distributions from prior knowledge. A couple of specific algorithms based on the MCMC theory are used to calibrate models in the previous literatures, such as Metropolis–Hasting (Sun et al., 2013), adaptive Metropolis algorithm (Hararuk et al., 2014), and multiple very fast simulated annealing (MVFSA) (Jackson et al., 2008). The MVFSA method is one to two orders of magnitude faster than the Metropolis–Hasting algorithm (Jackson et al., 2004). However, these methods only attempt to determine the most likely area of uncertain parameters and cannot directly give the best combination of uncertain parameters with a optimum metrics value. Moreover, the PDF heavily depends on the likelihood function assumed, which is usually difficult to determine for climate system model tuning problem.

Optimization algorithms can be used to search the maximum or minimum metrics value in a given parametric space. Severijns and Hazeleger (2005) calibrates parameters of radiation, clouds, and convection in Speedy model with the downhill simplex (Press et al., 1992; Nelder and Mead, 1965) to improve the radiation budget at the top of the atmosphere and at the surface, as well as the large scale circulation. The downhill simplex is a fast

convergence algorithm when the parametric space is not high-dimensional. However, it is a local optimization algorithm, not aiming to find the global optimal solution. Moreover, the algorithm has convergence issue when the simplex becomes ill-conditioned. Besides the downhill simplex, a few global optimization algorithms are introduced to tune uncertain parameters of climate system models, such as simulated stochastic approximation annealing (SSRR) (Yang et al., 2013), MVFSA (Yang et al., 2014), and multi-objective particle swarm optimization (MOPSO) (Gill et al., 2006). SSRR requires at least ten thousands of steps to get a stable solution (Liang et al., 2013), and MVFSA also requires thousands of steps to get a stable solution (Jackson et al., 2004). MOPSO needs dozens of individual cases in each iteration. All these global optimization algorithms require a large number of model runs and very high computational cost during the model tuning process.

Data assimilation method has been well addressed for state estimation, and can be a potential solution for parameter estimation. Aksoy et al. (2006) estimates the parameter uncertainty in a mesoscale model (Grell et al., 1994) using the Ensemble Kalman Filter (EnKF). Santitissadeekorn and Jones (2013) presents a two-step filtering for the joint state-parameter estimation with a combination method of particle filtering (PF) and EnKF. The EnKF and PF use an ensemble of model simulations to estimate the background error covariance, which approximate the traditional Kalman filter with a recurrence process (Evensen , 2003; Arulampalam et al., 2002). The accuracy of the error covariance relies on samples. In general, the larger the ensemble size, the more accurate the estimates are. The limitation of ensemble size for practice use and imperfect models make it difficult to select representative samples (Poterjoy et al., 2014). Moreover, same as the MOPSO method, they require a large number of model runs in each iteration with greatly increased computational cost.

Climate system model is a strongly nonlinear system, having a large number of uncertain parameters. As a result, the parametric space of a climate system model is high-dimensional, multi-modal, strongly nonlinear, unseparable. More seriously, one model run of a climate system model might require tens or even hundreds years of simulation to get scientifically meaningful results.

To overcome these challenges, we propose a "three-step" strategy to calibrate the uncertain parameters in climate system models effectively and efficiently. First, the Morris method (Morris, 1991; Campolongo et al., 2007), a global sensitivity analysis method, is chosen to eliminate the insensitive parameters by analyzing the main and interactive effects among parameters. Another global method by Sobol (Sobol, 2001) is used to validate the results of the Morris method. Second, a pre-processing of initial values of selected parameters is presented to accelerate the convergence of optimization algorithm and to resolve the issue of ill-conditioned problem. Finally, the downhill simplex algorithm is used to solve the optimization problem because of its low computational cost and fast convergence for low dimensional space. Taking into account the complex configuration and manipulation of model tuning, an automatic workflow is designed and implemented to make the calibration process more efficient. The method and workflow can be easily applied to GCMs to speed up model development process.

The paper is organized as follows. Section 2 introduces the proposed automatic workflow. Section 3 describes the details of the example model, reference data, and calibration metrics. The three-step calibration strategy is presented in Section 4. Section 5 evaluates the calibration results, followed by a summary in Section 6.

2 The end-to-end automatic calibration workflow

We design a software framework for the overall control of the tuning practice. This framework can automatically execute any part of our proposed "three-step" calibration strategy, determine the optimal parameters and produce its corresponding diagnostic results. It incorporates various tuning methods and facilitate model tuning process with minimal manual management. It effectively manages the dependence and calling sequences of various procedures, including parameter sampling, sensitivity analysis and initial value selection, model configuration and running, evaluation of model outputs using user provided metrics. Users only need to specify the model to tune, parameters to be tuned with their valid ranges, and the calibration method to use.

There are four main modules within the framework as shown in Fig. 1. The scheduler module manages model simulations with the capability for simultaneous runs. It also coordinates different tasks to reduce the contention and improve throughput. Simulation diagnosis and evaluation is included in a post-processing module. The preparation module contains various sensitivity analysis and sampling methods, such as Morris (Morris, 1991; Campolongo et al., 2007) and Sobol (Sobol, 2001) method, full factorial (FF) (Raktoe et al., 1981), Latin Hypercube (LH) (McKay et al., 1979), Morris one-at-a-time (MOAT) (Morris, 1991), and Central Composite Designs (CCD) (Hader and Park, 1978). The sensitivity analysis is able to eliminate the duplicated samples to reduce unnecessary model runs. A MCMC method based on adaptive Metropolis–Hastings algorithms is also provided to get the posterior distribution of uncertain parameters. The tuning algorithm module offers various local and global optimization algorithms including the downhill simplex, genetic algorithm, particle swarm optimization, differential evolution and simulated annealing. In addition, all the intermediate metrics and their corresponding parameters within the framework are stored in a MySQL database and can be used for posterior knowledge analysis. More importantly, the workflow is flexible and expandable for easy integration of other advanced algorithms as well as tools like the Problem Solving Environment for Uncertainty Analysis and Design Exploration (PSUADE) (Tong, 2005), Design Analysis Kit for Optimization and Terascale Applications (DAKOTA) (Eldred et al., 2007). Although, uncertainty quantification toolkits, such as PSUADE, DAKOTA, support various calibration and uncertainty analysis methods and pre-defined function interfaces, they cannot organize the above model tuning process as effectively as the proposed model tuning framework.

3 Model description and reference metrics

We use the Grid-point Atmospheric Model of IAP LASG version 2 (GAMIL2) as an example for the demonstration of the tuning workflow and our calibration strategy. GAMIL2 is the atmospheric component of the Flexible Global–Ocean–Atmosphere–Land System Model grid version 2 (FGOALS-g2), which participated in the CMIP5 program. The horizontal res-

olution is $2.8^\circ \times 2.8^\circ$, with 26 vertical levels. GAMIL2 uses a finite difference scheme that conserves mass and energy (Wang et al., 2004). A two-step shape-preserving advection scheme (Yu, 1994) is used for tracer advection. Compared to the pervious version, GAMIL2 has modifications in cloud-related processes (Li et al., 2013), such as the deep convection parameterization (Zhang and Mu, 2005), the convective cloud fraction (Xu and Krueger, 1991), the cloud microphysics (Morrison and Gettelman, 2008), and the stratiform fractional cloud condensation scheme (Zhang et al., 2003). More details are in Li et al. (2013). Empirical tunable parameters are selected from schemes of deep convection, shallow convection, and cloud fraction schemes (Table 1). Default parameter values are from the model configuration for CMIP5 experiments.

To save computational cost, atmosphere-only simulations are conducted for 5 years using prescribed seasonal climatology (no interannual variation) of SST and sea ice. Previous studies have shown 5 years of this type of simulation is enough to capture the basic characteristics of simulated mean climate states (Golaz et al., 2011; Lin et al., 2013). The goal of these simulations is not to determine their resemblance to observations, but to compare the results between the control simulation and various tuned simulations.

Model tuning results depend on the reference metrics used. For a simple justification, we use some conventional climate variables for the evaluation. Wind, humidity, and geopotential height are from the European Center for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA) – Interim reanalysis from 1989 to 2004 (Simmons et al., 2007). We use GPCP (Global Precipitation Climatology Project, Adler et al., 2003) for precipitation and ERBE (Earth Radiation Budget Experiment, Barkstrom, 1984) for radiative fields. All observational and reanalysis data are gridded to the same grid as GAMIL2 before the comparison. Note that the reference metrics can be extended depending on the model performance requirement.

The reference metrics, including various variables in Table 2, is used to quantitatively evaluate the performance of overall simulation skills (Murphy et al., 2004; Gleckler et al., 2008; Reichler and Kim, 2008). The calibration RMSE is defined as the spatial standard deviation (SD) of the model simulation against observations/re-analysis, as in Eq. (1) (Taylor,

2001; Yang et al., 2013). For an easy comparison, we normalize the RMSE of each simulation output by that of the control simulation using default parameter values. We introduce an improvement index to evaluate the tuning results, which weight each variable equally and compute the average normalized RMSE. The index indicates an overall improvement of the performance of the tuned simulation relative to the control simulation based on a number of model outputs (Table 2). If the index is less than 1, it means the tuned simulation gets better performance than the control run. The smaller this value, the better improvement is.

$$(\sigma_{\mathsf{m}}^F)^2 = \sum_{i=1}^{I} w(i)(x_{\mathsf{m}}^F(i) - x_{\mathsf{o}}^F(i))^2 \tag{1}$$

$$(\sigma_r^F)^2 = \sum_{i=1}^{I} w(i)(x_r^F(i) - x_{\mathsf{o}}^F(i))^2 \tag{2}$$

$$\chi^2 = \frac{1}{N^F} \sum_{F=1}^{N^F} (\frac{\sigma_{\mathsf{m}}^F}{\sigma_r^F})^2 \tag{3}$$

$x_{\mathsf{m}}^F(i)$ is the model outputs, and $x_{\mathsf{o}}^F(i)$ is the corresponding observation or reanalysis data. $x_r^F(i)$ is model outputs from the control simulation using the default values for the parameters in Table 1. w is the weight due to the different grid area on a regular latitude longitude grids on the sphere. I is the total grid number in model. N^F is the number of the chosen variables.

4 Method

4.1 Parameters tuning with global and local optimization methods

Parameter tuning for a climate system model is to solve a global optimization problem in theory. As the well-known global optimization algorithms, traditional evolutionary algorithms, such as genetic algorithm (Goldberg et al., 1989), differential evolutionary (DE) (Storn and

Price, 1995), and particle swarm optimization (PSO) (Kennedy, 2010), can approach the global optimal solution but generally require high computational cost. This is because these algorithms are designed following biological evolution of survival of the fittest. In contrast, the local algorithms utilize the greedy strategy, and thus may stick at a locally optimal solution after convergence. The advantage of local algorithms is the low computational cost due to relatively less samples required. In this sense, the local optimization algorithms are the viable options considering their significantly reduced computational cost.

We choose the downhill simplex method for climate model tuning considering its relatively low computation cost. The downhill simplex method searches the optimal solution by changing the shape of a simplex, which represents the optimal direction and step length. A simplex is a geometry, consisting of $N+1$ vertexes and their interconnecting edges, where *N* is the number of calibration parameters. One vertex stands for a pair of a set of parameters and their improvement index as defined in Eq. (3). The new vertex is determined by expanding or shrinking the vertex with the highest metrics value, leading to a new simplex (Press et al., 1992; Nelder and Mead, 1965).

Two performance criteria are used to evaluate the effectiveness and efficiency of the optimization algorithms in this study. Selection of optimization algorithms for parameter calibration of climate system models is a balance between model improvement (effectiveness) and computational cost (efficiency). In this study, model improvement is measured by an index defined in Eq. (3). The lower of this value is, the better model tuning is. Computational cost is measured by "core-hours", standing for the computational efficiency. It is computed by $(N_{step}) \times (N_{size}) \times$ (the number of processes of a single model run) $\times$ (hours used for a single 5-year model run). N_{step} is the total numbers of iterations of optimization algorithms for convergence. N_{size} is the number of model runs during each iteration, and it is 1 for the downhill simplex method. In GAMIL2 case, each model run takes 6 hours using 30 processes.

According to tuning GAMIL2, two global methods, PSO and DE, give better tuning effectiveness than the downhill simplex method, but their computational costs are approximately 4 and 5 times of the downhill simplex method, respectively (Table 3).

To improve the effectiveness of the downhill simplex method, we propose two important steps to significantly improve its performance. In the first step, the number of tuning parameters is reduced by eliminating the insensitive parameters; In the second step, fast convergence is achieved by pre-selecting proper initial values for the parameters before using the downhill simplex method.

4.2 Parameter sensitivity analysis

The number of uncertain parameters in physical parameterizations of a climate system model is quite large. Most optimization algorithms, such as PSO, the downhill simplex method, and the simulated annealing algorithm (Van Laarhoven and Aarts, 1987), are ineffective in high dimensional problems. Iterations for convergence will increase exponentially with the number of tuning parameters. In addition, climate models generally need a long simulation to have meaningful results. Therefore, high dimensional parameter tuning problem suffers from extremely high computational cost. It is necessary to reduce the parameter dimension before the optimization.

Parameter sensitivity analysis can be divided into local and global methods (Gan et al., 2014). The local method determines the sensitivity of a single parameter by perturbing one parameter with all other parameters fixed. Consequently, it does not consider the combined sensitivity of multiple parameters. On the other hand, the global method perturbs all the parameters to explore the sensitivity of the whole parametric space. In this study, the Morris (Morris, 1991; Campolongo et al., 2007), a global method, is used to screen out the sensitive parameters. Another global method (Sobol, 2001), is used to validate the results of the Morris method.

The Morris method, based on the MOAT sampling strategy, reduces the number of samples required by other global sensitivity methods (Li et al., 2013). Note that a sample is a set of all parameters, not just one parameter. The method is described briefly here, and more details can be found in Morris (1991). Assume we have k parameters, relative to a random sample $S_1 = \{x_1, x_2, ..., x_k\}$, another sample $S_2 = \{x_1, x_2, ..., x_i + \Delta_i, ..., x_k\}$ can be constructed by perturbing the i_{th} parameter by Δ_i, where Δ_i is a perturbation of this

parameter. The elementary effect of the i_{th} parameter x_i is defined as:

$$d_i = \frac{f(S_2) - f(S_1)}{\Delta_i} \tag{4}$$

where f stands for the improvement index as defined in Eq. (3). A third sample S_3 = $\{x_1, x_2, ..., x_i + \Delta_i, ..., x_j + \Delta_j, ..., x_k\}$ can be generated by perturbing another parameter, where j is not i. In so doing k times, we will get k+1 samples $\{S_1, S_2, ..., S_{k+1}\}$, and k elementary effects $\{d_1, d_2, ..., d_k\}$ after perturbing all the parameters. The vector of $\{S_1, S_2, \ldots, S_{k+1}\}$ is called a trajectory. This procedure is repeated for r iterations and finally we get r trajectories. The starting point of any trajectory is selected randomly as well as the ordering of the parameters to perturb and the Δ for each perturbation in one trajectory. In practice, a number of 10 to 50 trajectories are enough to determine the feasible sensitivity of parameters (Gan et al., 2014; Morris, 1991). In this study, we have a total of 7 parameters and 80 simulations are conducted.

We define D = $\{D_i(t)\}$, where t is the t_{th} trajectory, and i is the i_{th} elementary effect of the parameter x_i. μ_i, the mean of $|d_i|$, and σ_i, the standard deviation of d_i, are used to measure the parameter sensitivity, defined as:

$$\mu_i = \sum_{t=1}^{r} \frac{|d_i(t)|}{r} \tag{5}$$

$$\sigma_i = \sum_{t=1}^{r} \sqrt{(d_i(t) - \mu_i)^2 / r} \tag{6}$$

μ_i estimates the effect of x_i on the model improvement index as defined in Eq. (3), while σ_i assesses interactive effect of x_i with other parameters. Those parameters with large μ_i and σ_i are the sensitive parameters. The Morris method results are shown in Fig. 2.

The parameter elimination step is critical for the final result of model tuning. To validate the results obtained by the Morris method, we compare the results with a benchmark method (Sobol, 2001). Based on variance decomposition, the Sobol method requires more

samples than the Morris method, leading to a higher computation cost. The variance of the model output can be decomposed as Eq. (7), where *n* is the number of parameters, and V_i is the variance of the ith parameter, and V_{ij} is the variance of the interactive effect between the ith and jth parameters, and so on. The total sensitivity effect of ith parameter can be presented as Eq. (8), where V_{-i} is the total variance except for the x_i parameter. The Sobol results are shown in Fig. 3. The screened out parameters are the same as those of the Morris.

$$V = \sum_{i=1}^{n} V_i + \sum_{1 \le i < j \le n} V_{ij} + \ldots + V_{1,2,\ldots,n} \tag{7}$$

$$S_{T_i} = 1 - \frac{V_{-i}}{V} \tag{8}$$

4.3 Proper initial value selection for the downhill simplex method

The downhill simplex method is a local optimization algorithm and its convergence performance strongly depends on the quality of the initial values. We need to find the parameters with the smaller metrics around the final solution. Moreover, we have to finish the searching as fast as possible with minimal overhead. For these two objectives, a hierarchical sampling strategy based on the single parameter perturbation (SPP) sample method is used. The SPP is similar to local sensitivity methods, in which only one parameter is perturbed at one time with other parameters fixed. The perturbation samples are uniformly distributed across parametric space. First, the improvement index as defined in Eq. (3) of each parameter sample is computed. The distance is defined as the difference between the improvement indexes using two adjacent samples, i.e., the model response measured by certain percentage change of one parameter. We call this step the first level sampling. The specific perturbation size for one parameter can be set based on user experience. In our implementation, user needs to set the number of samples. For the first level sampling, we can use a larger perturbation size to reduce computational cost. If the distance between two adjacent samples is greater than a predefined threshold, more SPP samples between the

previous two adjacent samples are conducted. And this is called the second level sampling. Finally, k+1 samples with the best improvement index value are chosen as the candidate initial values for the optimization method. With this hierarchical sampling strategy, we can determine the local parametric space for final solution and can accelerate the convergence of the following downhill simplex method. This procedure is described in Algorithm 1. It is easy to implement and has lower overhead compared to other complex adaptive sampling methods.

At the same time, inappropriate initial values may lead to ill-conditioned simplex geometry, which can be found in model tuning process. One issue we meet is that some vertexes in the downhill simplex optimization may have the same values for one or more parameters. As a result, these parameters remain invariant during the optimization and this may degrade the quality of final solution as well as the convergence speed. A simplex checking is conducted to keep as many different values of parameters as possible during the process of looking for initial values. Well-conditioned simplex geometry will increase the parameter freedom for optimization. In our implementation (Algorithm 1), the vertex leading to the ill-conditioned simplex is replaced by another parameter sample which gives another minimum improvement index value.

These methods mentioned above are summarized as the initial value pre-processing of the downhill simplex algorithm. Sometimes, the samples used during the initial value selection are the same as those in the parameter sensitivity analysis step. In this case, one model run can be used in both steps to further reduce the computational cost.

4.4 Evaluation of the proposed strategy

Effectiveness and efficiency of the three traditional algorithms are compared in Table 3. “Downhill_1_step” represents the original downhill simplex method, which is one of the most widely-used local optimization algorithms and has been successfully used in Speedy model (Severijns and Hazeleger, 2005). PSO and DE are the most widely-used global optimization algorithms and easy-to-use. Although “Downhill_1_step” achieves slightly worse

improvement compared to the two global optimization methods (Table 3), its computation cost is much less (only 20% and 28% of DE and PSO respectively).

Two extra steps are included before the original downhill simplex method to overcome its limited effectiveness on model performance improvement. The "Downhill_2_steps" method includes an initial value pre-processing step before the downhill simplex method. And the "Downhill_3_steps" method further introduces another step to eliminate insensitive parameters for tuning by sensitivity analysis. The two steps bring in additional overhead, 80 samples for the parameter sensitivity analysis with the Morris method, and 25 samples for the initial value pre-processing. Table 3 and 4 show that the proposed "Downhill_3_steps" achieves the best effectiveness, improving the model's overall performance by 9%. It overcomes the inherent ineffectiveness of the original downhill simplex method with much lower computational cost than global methods.

5 Analysis of model optimal results

This section compares the default simulation and the tuned simulation by three-step method with a focus on the cloud and TOA radiation changes. Table 1 shows the values of the four pairs of sensitive parameters between the control (labeled as CNTL) and optimized simulation (labeled as EXP). Significant change is found for c0, which represents the autoconversion coefficient in the deep convection scheme, and rhminh, which represents the threshold relative humidity for high cloud appearance. The other two parameters have negligible change of the values before and after the tuning and thus it is expected their impacts on model performance will be accordingly small.

The overall improvement after the tuning from the control simulation can be found in the Taylor diagram (Fig. 4), with improvement for almost all the variables, especially for the meridional winds and mid-tropospheric (400 hPa) humidity. Improvements for other variables are relatively small. The change in terms of the RMSE factor over the globe and three regions (tropics, SH mid- and high-latitude and NH mid- and high-latitude) are shown in Fig. 5. First, radiative fields and moisture are improved over all the four areas. By contrast,

wind and temperature field changes are more diverse among different areas. For example, temperatures over the tropics become worse compared to the control run. There is an overall improvement in the SH mid- and high-latitude for all variables except for the 200 hPa temperature. Winds and precipitation in the NH mid- and high-latitude become slightly worse in the tuned simulation. Such changes are kind of intriguing and we attempt to relate these changes to the two parameters significantly tuned.

With reduced RH threshold for high cloud (from 0.78 in CNTL to 0.63 in EXP, Table 1), the stratiform condensation rate increases and the atmospheric humidity decreases (Zhang et al., 2003). In addition, with increased auto-conversion coefficient in the deep convection, less condensate is detrained to the environment. As a result, mid- and upper-troposphere is overall drier, especially over the tropics where deep convection dominates the vertical moisture transport (Fig. 6c). Although the mid- and upper-troposphere become drier over the tropics, reduced RH threshold for high cloud makes clouds easier to be present. Consequently, middle and high clouds increase over the globe, especially over the mid- and high-latitudes with the largest increase up to 4–5 % (Fig. 6f). In the tropics, due to the drier tendency induced by the reduced detrainment, high cloud increase is relatively small (2–3 %) compared to the mid- and high-latitudes. On the contrary, low cloud below 800 hPa decreases by 1–2 % over the mid- and high-latitudes with slightly decreased RH (Fig. 6) because of the negligible change of RH threshold for low cloud (Table 1). Overall, the combined effects of all relevant parameterizations lead to the changes of atmospheric humidity and cloud fraction.

Changes in moisture and cloud fields impact radiative fields. With reference to ERBE, TOA outgoing longwave radiation (OLR) is improved in the mid-latitudes for EXP, but it is degraded over the tropics (Fig. 7a). Compared with the CNTL, middle and high cloud significantly increase in the EXP (Fig. 6). Consequently, it enhances the blocking effect on the longwave upward flux at TOA (FLUT), reducing the FLUT in mid-latitudes of the southern and Northern Hemisphere (Fig. 7a). Clear sky OLR increases for the EXP and this is due to the drier upper troposphere in the EXP (Fig. 6). The decrease in the atmospheric water vapor reduces the greenhouse effect. Therefore, it emits more outgoing longwave radiation

and reduces the negative bias of clear sky long wave upward flux at TOA (FLUTC, Fig. 7b). Longwave cloud forcing (LWCF) in the middle and high latitudes is improved due to the improvement of FLUT in these areas (Fig. 7c), but improvement in the tropics is negligible due to the cancellation between the FLUT and FLUTC.

TOA clear sky shortwave are the same between the control and the tuned simulation since both simulations have the same surface albedo. With increased clouds, the tuned simulation has smaller TOA shortwave absorbed than the control. Compared with ERBE, the tuned simulation has better TOA shortwave absorbed in the mid- and high-latitudes, but it slightly degrades over the tropics.

6 Conclusions

An effective and efficient three-step method for GCM physical parameter tuning is proposed. Compared with conventional methods, a parameter sensitivity analysis step and a proper initial value selection step are introduced before the low cost downhill simplex method. This effectively reduces the computational cost with an overall good performance. In addition, an automatic parameter calibration workflow is designed and implemented to enhance operational efficiency and support different uncertainty quantification analysis and calibration strategies. Evaluation of the method and workflow by calibrating GAMIL2 model indicates the three-step method outperforms the two global optimization methods (PSO and DE) in both effectiveness and efficiency. A better trade-off between accuracy and computational cost is achieved compared with the two-step method and the original downhill simplex method. The optimal results of the three-step method demonstrate that most of the variables are improved compared with the control simulation, especially for the radiation related ones. The mechanism analysis is conducted to explain why these radiation related variables have an overall improvement. In future work, more analyses are needed to better understand the model behavior along with the physical parameter changes.

Algorithm 1 Preprocessing the initial values of Downhill Simplex Algorithm.

```
//Single parameter perturbation sample(SPP)
N=number_of_parameters
sampling_sets={}
for each parameter P_i of N parameters do
    sampling_sets+=SPP(P_i_range, number_of_samples)
    //refine sample in the sensitivity range if needed
    if metrics of the the adjacent same parameter sampling points >= sensitivity_threhold then
        sampling_sets+=SPP(P_i_adjacent_parameter_range, refine_ number_of_factors)
    end if
end for

//Initial vertexes with parameters of the N + 1 minimum metrics
for each initial V_i of N + 1 vertexes do
    //get the parameters of the ith minimum metrics
    candidate_init_sets += min(i, sampling_sets)
end for

//make sure the initial simplex geometry is well-conditioned
while one parameter k have the same values in the N + 1 sets do
    j = 1
    //remove the parameter set with the worst metrics from candidate_init_sets
    remove_parameter_set(the parameter set with worse metrics, candidate_init_sets)
    //get the parameters of the N + 1 + jth minimum metrics
    candidate_init_sets += min(N + 1 + j, sampling_sets)
    j+ = 1
end while
```

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., Rudolf, B., Schneider, U., Curtis, S., Bolvin, D., Gruber, A., Susskind, J., Arkin, P., and Nelkin, E.: The version-2

global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present), J. Hydrometeorol., 4, 1147–1167, 2003.
Aksoy, A., Zhang, F., and Nielsen-Gammon, J. W.: Ensemble-based simultaneous state and parameter estimation with MM5, Geophys. Res. Lett., 33, L12801, doi:10.1029/2006GL026186, 2006.
Allen, M. R., Stott, P. A., Mitchell, J. F., Schnur, R., and Delworth, T. L.: Quantifying the uncertainty in forecasts of anthropogenic climate change, Nature, 407, 617–620, 2000.
Arulampalam, M., Maskell, S., Gordon, N., and Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, Signal Processing, IEEE Transactions on, 50, 174–188, 2002.
Bardenet, R., Brendel, M., Kégl, B., and Sebag, M.: Collaborative hyperparameter tuning, in: Proceedings of the 30th International Conference on Machine Learning (ICML-13), 16–21 June 2013, Atlanta, Georgia, USA, 199–207, 2013.
Barkstrom, B. R.: The earth radiation budget experiment (ERBE), B. Am. Meteorol. Soc., 65, 1170–1185, 1984.
Beven, K. and Binley, A.: The future of distributed models: model calibration and uncertainty prediction, Hydrol. Process., 6, 279–298, 1992.
Cameron, D., Beven, K. J., Tawn, J., Blazkova, S., and Naden, P.: Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty), J. Hydrol., 219, 169–187, 1999.
Campolongo, F., Cariboni, J., and Saltelli, A.: An effective screening design for sensitivity analysis of large models, Environ. Modell. Softw., 22, 1509–1518, 2007.
Chen, T.-Y., Wei, W.-J., and Tsai, J.-C.: Optimum design of headstocks of precision lathes, Int. J. Mach. Tool. Manu., 39, 1961–1977, 1999.
Eldred, M., Agarwal, H., Perez, V., Wojtkiewicz Jr., S., and Renaud, J.: Investigation of reliability method formulations in DAKOTA/UQ, Struct. Infrastruct. E., 3, 199–213, 2007.
Elkinton, C. N., Manwell, J. F., and McGowan, J. G.: Algorithms for offshore wind farm layout optimization, Wind Engineering, 32, 67–84, 2008.
Evensen, G.: The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean dynamics, 53, 343–367, 2003.
Gan, Y., Duan, Q., Gong, W., Tong, C., Sun, Y., Chu, W., Ye, A., Miao, C., and Di, Z.: A comprehensive evaluation of various sensitivity analysis methods: a case study with a hydrological model, Environ. Modell. Softw., 51, 269–285, 2014.

Gilks, W. R.: Markov Chain Monte Carloin Practice, Chapman and Hall/CRC, London, United Kingdom, 1995.

Gill, M. K., Kaheil, Y. H., Khalil, A., McKee, M., and Bastidas, L.: Multiobjective particle swarm optimization for parameter estimation in hydrology, Water Resour. Res., 42, W07417, doi:10.1029/2005WR004528, 2006.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res.-Atmos., 113, D6, doi:10.1029/2007JD008972, 2008.

Golaz, J., Salzmann, M., Donner, L. J., Horowitz, L. W., Ming, Y. and Zhao, M.: Sensitivity of the aerosol indirect effect to subgrid variability in the cloud parameterization of the GFDL atmosphere general circulation model AM3, J. Climate, 24, 3145–3160, 2011.

Goldberg, D. E., Korb, B., and Deb, K.: Messy genetic algorithms: motivation, analysis, and first results, Complex systems, 3, 493–530, 1989.

Grell, G. A., Dudhia, J., and Stauffer, D.: A description of the fifthgeneration Penn State/NCAR Mesoscale Model (MM5), NCAR Tech. Note NCAR/TN-3981STR, 122 pp., 1994.

Hack, J. J., Boville, B., Kiehl, J., Rasch, P., and Williamson, D.: Climate statistics from the National Center for Atmospheric Research Community Climate Model CCM2, J. Geophys. Res.-Atmos., 99, 20785–20813, 1994.

Hader, R. and Park, S. H.: Slope-rotatable central composite designs, Technometrics, 20, 413–417, 1978.

Hakkarainen, J., Ilin, A., Solonen, A., Laine, M., Haario, H., Tamminen, J., Oja, E., and Järvinen, H.: On closure parameter estimation in chaotic systems, Nonlin. Processes Geophys., 19, 127–143, doi:10.5194/npg-19-127-2012, 2012.

Hararuk, O., Xia, J., and Luo, Y.: Evaluation and improvement of a global land model against soil carbon data using a Bayesian Markov chain Monte Carlo method, J. Geophys. Res.-Biogeo., 119, 403–417, 2014.

Hegerty, B., Hung, C.-C., and Kasprak, K.: A comparative study on differential evolution and genetic algorithms for some combinatorial problems, in: Proceedings of 8th Mexican International Conference on Artificial Intelligence, 9–13 November 2009, Guanajuato, Mexico, 88, 2009.

Jackson, C., Sen, M. K., and Stoffa, P. L.: An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions, J. Climate, 17, 2828–2841, 2004.

Jackson, C. S., Sen, M. K., Huerta, G., Deng, Y., and Bowman, K. P.: Error reduction and convergence in climate prediction, J. Climate, 21, 6698–6709, 2008.

Jakumeit, J., Herdy, M., and Nitsche, M.: Parameter optimization of the sheet metal forming process using an iterative parallel Kriging algorithm, Struct. Multidiscip. O., 29, 498–507, 2005.

Kennedy, J.: Particle swarm optimization, in: Encyclopedia of Machine Learning, Springer, New York, USA, 760–766, 2010.

Li, L., Wang, B., Dong, L., Liu, L., Shen, S., Hu, N., Sun, W., Wang, Y., Huang, W., Shi, X., Pu, Y., and Yang, G.: Evaluation of grid-point atmospheric model of IAP LASG version 2 (GAMIL2), Adv. Atmos. Sci., 30, 855–867, 2013.

Li, J., Duan, Q. Y., Gong, W., Ye, A., Dai, Y., Miao, C., Di, Z., Tong, C. and Sun, Y.: Assessing parameter importance of the Common Land Model based on qualitative and quantitative sensitivity analysis, Hydrol. Earth Syst. Sci, 17, 3279–3293, 2013.

Liang, F., Cheng, Y., and Lin, G.: Simulated stochastic approximation annealing for global optimization with a square-root cooling schedule, J. Am. Stat. Assoc., 109, 847–863, 2013.

Lin, Y. L., Zhao, M., Ming, Y., Golaz, J., Donner, L. J., Klein, S. A., Ramaswamy, V. and Xie, S.: Precipitation partitioning, tropical clouds, and intraseasonal variability in GFDL AM2, J. Climate, 26, 5453–5466, 2013.

McKay, M. D., Beckman, R. J., and Conover, W. J.: Comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21, 239–245, 1979.

Morris, M. D.: Factorial sampling plans for preliminary computational experiments, Technometrics, 33, 161–174, 1991.

Morrison, H. and Gettelman, A.: A new two-moment bulk stratiform cloud microphysics scheme in the Community Atmosphere Model, version 3 (CAM3). Part I: Description and numerical tests, J. Climate, 21, 3642–3659, 2008.

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., and Stainforth, D. A.: Quantification of modelling uncertainties in a large ensemble of climate change simulations, Nature, 430, 768–772, 2004.

Nelder, J. A. and Mead, R.: A simplex method for function minimization, Computer J., 7, 308–313, 1965.

Poterjoy, J., Zhang, F., and Weng, Y.: The effects of sampling errors on the EnKF assimilation of inner-core hurricane observations, Mon. Weather Rev., 142, 1609–1630, 2014.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B.: Numerical Recipes in Fortran, Cambridge Univ. Press, Cambridge, 70 pp., 1992.

Raktoe, B. L., Hedayat, A., and Federer, W. T.: Factorial Designs, John Wiley & Sons, Hoboken, New Jersey, USA, 1981.
Reichler, T. and Kim, J.: How well do coupled models simulate today's climate?, B. Am. Meteorol. Soc., 89, 303–311, 2008.
Santitissadeekorn, N. and Jones, C.: Two-stage filtering for joint state-parameter estimation, Mon. Weather Rev., accepted, doi:10.1175/MWR-D-14-00176.1, 2013.
Severijns, C. and Hazeleger, W.: Optimizing parameters in an atmospheric general circulation model, J. Climate, 18, 3527–3535, 2005.
Shi, Y. and Eberhart, R. C.: Empirical study of particle swarm optimization, in: Proceedings of the 1999 Congress on Evolutionary Computation, CEC 99, vol. 3, IEEE, 6–9 July 1999, Washington, DC, USA, 1945–1950, 1999.
Simmons, A., Uppala, S., Dee, D., and Kobayashi, S.: ERA-interim: new ECMWF reanalysis products from 1989 onwards, ECMWF Newsl., 110, 25–35, 2007.
Sobol, I. M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Math. Comput. Simulat., 55, 271–280, 2001.
Storn, R. and Price, K.: Differential Evolution – a Simple and Efficient Adaptive Scheme for Global Optimization over Continuous Spaces, ICSI Berkeley, Berkeley, California, USA, 1995.
Sun, Y., Hou, Z., Huang, M., Tian, F., and Ruby Leung, L.: Inverse modeling of hydrologic parameters using surface flux and runoff observations in the Community Land Model, Hydrol. Earth Syst. Sci., 17, 4995–5011, doi:10.5194/hess-17-4995-2013, 2013.
Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res.-Atmos., 106, 7183–7192, 2001.
Tong, C.: PSUADE User's Manual, Lawrence Livermore National Laboratory (LLNL), Livermore, CA, 109 pp., 2005.
Van Laarhoven, P. J. and Aarts, E. H.: Simulated Annealing, Springer, Dordrecht, Netherlands, 1987.
Wang, B., Wan, H., Ji, Z., Zhang, X., Yu, R., Yu, Y., and Liu, H.: Design of a new dynamical core for global atmospheric models based on some efficient numerical methods, S. China Ser. A, 47, 4–21, 2004.
Warren, S. G. and Schneider, S. H.: Seasonal simulation as a test for uncertainties in the parameterizations of a Budyko-Sellers zonal climate model, J. Atmos. Sci., 36, 1377–1391, 1979.
Williams, P. D.: Modelling climate change: the role of unresolved processes, Philos. T. R. Soc. A, 363, 2931–2946, 2005.

Xu, K.-M. and Krueger, S. K.: Evaluation of cloudiness parameterizations using a cumulus ensemble model, Mon. Weather Rev., 119, 342–367, 1991.

Yang, B., Qian, Y., Lin, G., Leung, L. R., Rasch, P. J., Zhang, G. J., McFarlane, S. A., Zhao, C., Zhang, Y., Wang, H., Wang, M., and Liu, X.: Uncertainty quantification and parameter tuning in the CAM5 Zhang-McFarlane convection scheme and impact of improved convection on the global circulation and climate, J. Geophys. Res.-Atmos., 118, 395–415, 2013.

Yang, B., Zhang, Y., Qian, Y., Huang, A., and Yan, H.: Calibration of a convective parameterization scheme in the WRF model and its impact on the simulation of East Asian summer monsoon precipitation, Clim. Dynam., 1, 1–24, 2014.

Yu, R.: A two-step shape-preserving advection scheme, Adv. Atmos. Sci., 11, 479–490, 1994.

Zhang, G. J. and Mu, M.: Effects of modifications to the Zhang-McFarlane convection parameterization on the simulation of the tropical precipitation in the National Center for Atmospheric Research Community Climate Model, version 3, J. Geophys. Res.-Atmos., 110, D09109, 10.1029/2004JD005617, 2005.

Zhang, M., Lin, W., Bretherton, C. S., Hack, J. J. and Rasch, P. J.: A modified formulation of fractional stratiform condensation rate in the NCAR Community Atmospheric Model (CAM2), J. Geophys. Res.-Atmos., 108(D1),ACL–10, 2003.

Table 1. A summary of parameters to be tuned in GAMIL2. The default and the final tuned optimum value are also shown. The valid range of each parameter is also included. Note that only four sensitive parameters are tuned and have optimum values.

Parameter	Description	Default	Range	Optimal
c0	rain water autoconversion coefficient for deep convection	3.0×10^{-4}	$1. \times 10^{-4}$–5.4×10^{-3}	5.427294×10^{-4}
ke	evaporation efficiency for deep convection	7.5×10^{-6}	5×10^{-7}–5×10^{-5}	–
capelmt	threshold value for cape for deep convection	80	20–200	–
rhminl	threshold RH for low clouds	0.915	0.8–0.95	0.917661
rhminh	threshold RH for high clouds	0.78	0.6–0.9	0.6289215
c0_shc	rain water autoconversion coefficient for shallow convection	5×10^{-5}	3×10^{-5}–2×10^{-4}	–
cmftau	characteristic adjustment time scale of shallow cape	7200	900–14 400	7198.048

Table 2. Atmospheric fields included in the evaluation metrics and their sources.

Variable	Observation	Variable	Observation
Meridional wind at 850 hPa	ECMWF	Geopotential Z at 500 hPa	ECMWF
Meridional wind at 200 hPa	ECMWF	Total precipitation rate	GPCP
Zonal wind at 850 hPa	ECMWF	Long-wave cloud forcing	ERBE
Zonal wind at 200 hPa	ECMWF	Short-wave cloud forcing	ERBE
Temperature at 850 hPa	ECMWF	Long-wave upward flux at TOA	ERBE
Temperature at 200 hPa	ECMWF	Clearsky long-wave upward flux at TOA	ERBE
Specific Humidity at 850 hPa	ECMWF	Short-wave net flux at TOA	ERBE
Specific Humidity at 400 hPa	ECMWF	Clearsky short-wave net flux at TOA	ERBE

Table 3. Effectiveness and efficiency comparison between the original downhill simplex method and the two global methods. N_{step} is the total numbers of calibrating iteration for convergence. N_{size} is is the number of model runs during each iteration. Core-hours is computed by $N_{\text{step}} \times N_{\text{size}} \times$ {the number processes of a single model run} $\times$ {hours used for a single 5-year model run }. In GAMIL2 case, each model run takes 6 h and using 30 processes.

	Improvement index	N_{step}	N_{size}	Core-hours
Downhill_1_step	0.9585	80	1	14 400
PSO	0.9115	24	12	51 840
DE	0.9421	33	12	71 280

Table 4. The same as Table 3, but showing the comparison among the three downhill simplex methods.

	Improvement index	N_{step}	N_{size}	Core hours
Downhill_1_step	0.9585	80	1	14 400
Downhill_2_steps	0.9257	$25 + 34$	1	10 620
Downhill_3_steps	0.9099	$80 + 25 + 50$	1	27 900

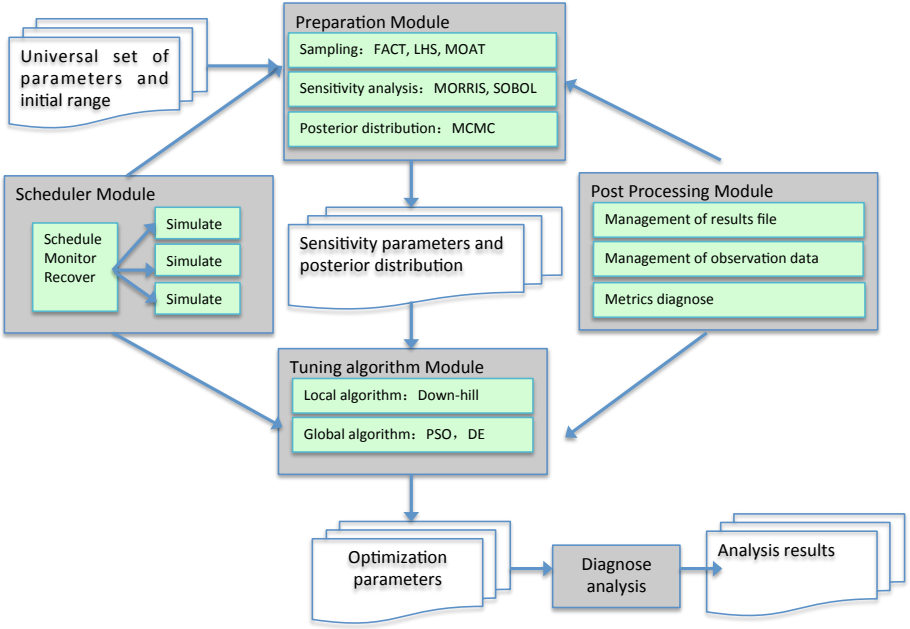


Figure 1. The structure of the automatic calibration workflow. The input of the workflow is the parameters of interest and their initial value ranges. The output is the optimal parameters and its corresponding diagnostic results after calibration. The preparation module provides the parameter sensitivity analysis. The tuning algorithm module offers local and global optimization algorithms including downhill simplex, genetic algorithm, particle swarm optimization, differential evolution and simulated annealing. The scheduler module schedules as many as cases to run simultaneously and coordinates different tasks over parallel system. The post-processing module is responsible for metrics diagnostics, re-analysis and observational data management.

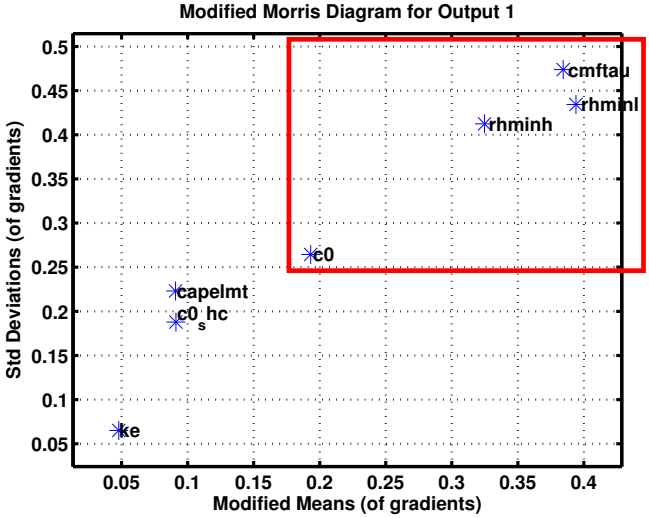


Figure 2. Scatter diagram showing the parameter sensitivity using the Morris sensitivity analysis. The x axis stands for the main effect sensitivity of a single parameter. The y axis stands for the interactive effect sensitivity among multi-parameters. In GAMIL2, c0, rhminl, rhminh, and cmftau have high sensitivity and ke, c0_shc, and capelmt have low sensitivity.

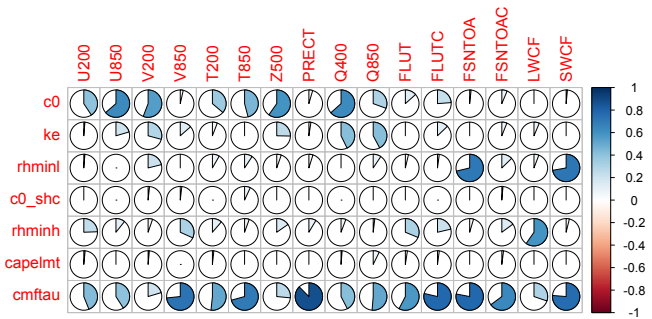


Figure 3. Sensitivity analysis results from the Sobel method. The total sensitivity in Eq. (8) is denoted by the size of color area. The total sensitivities of ke, c0_shc, and capelmt are less than 0.5 in terms of each variable.

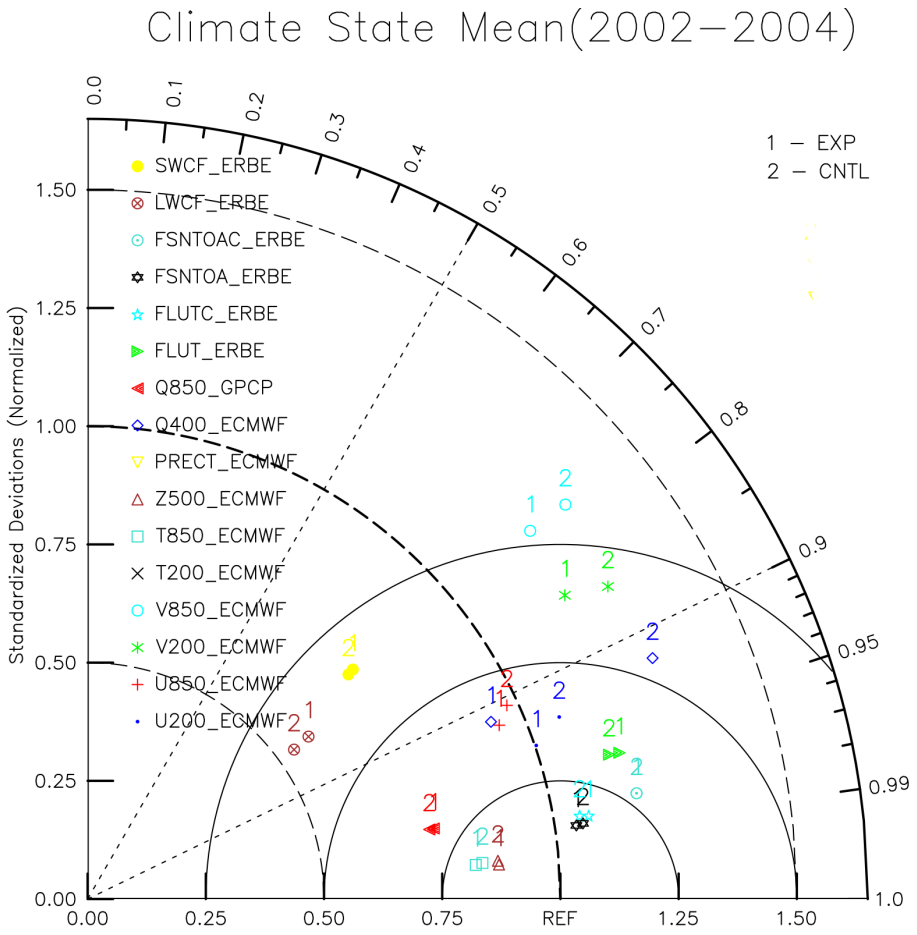


Figure 4. Taylor diagram of the climate mean state of each output variable from 2002 to 2004 of EXP and CNTL.

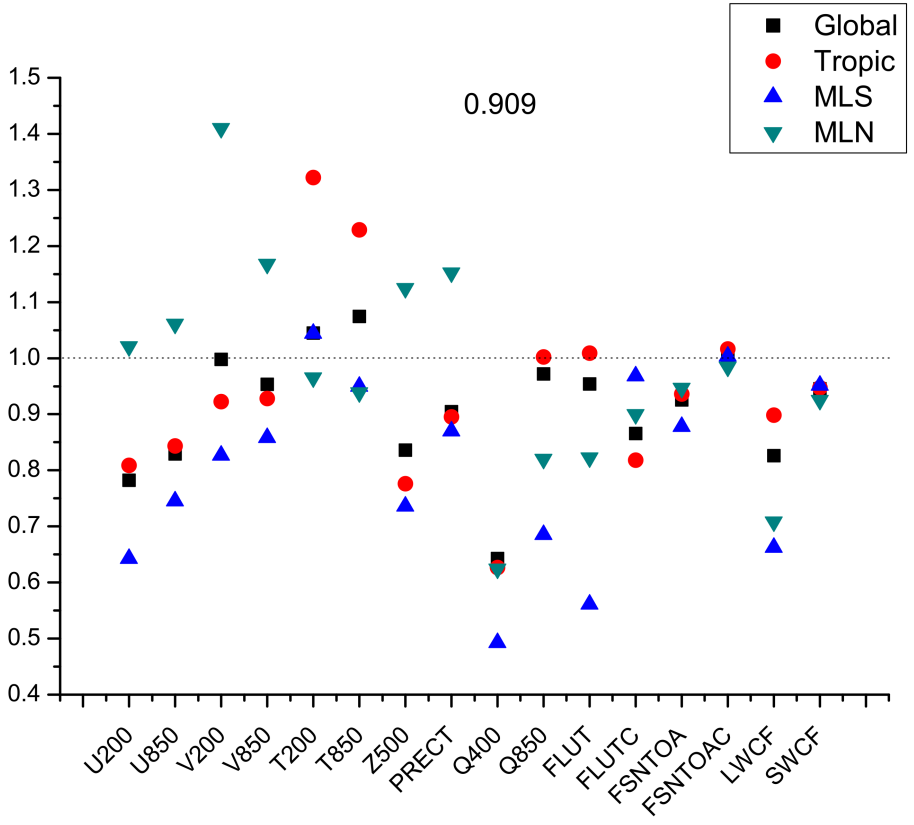


Figure 5. Improvement indices over the global, tropical and mid-high latitudes of northern and southern hemisphere (MLN and MLS) for each variable of the EXP simulation.

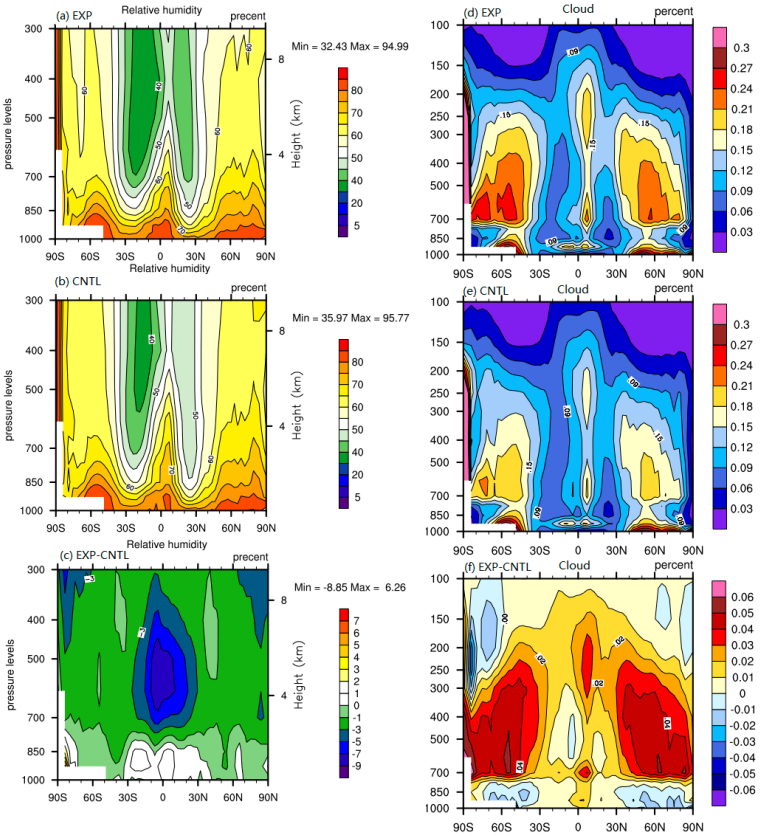


Figure 6. Pressure–latitude distributions of relative humidity and cloud fraction of EXP **(a, d)**, CNTL **(b, e)**, EXP-CNTL **(c, f)**.

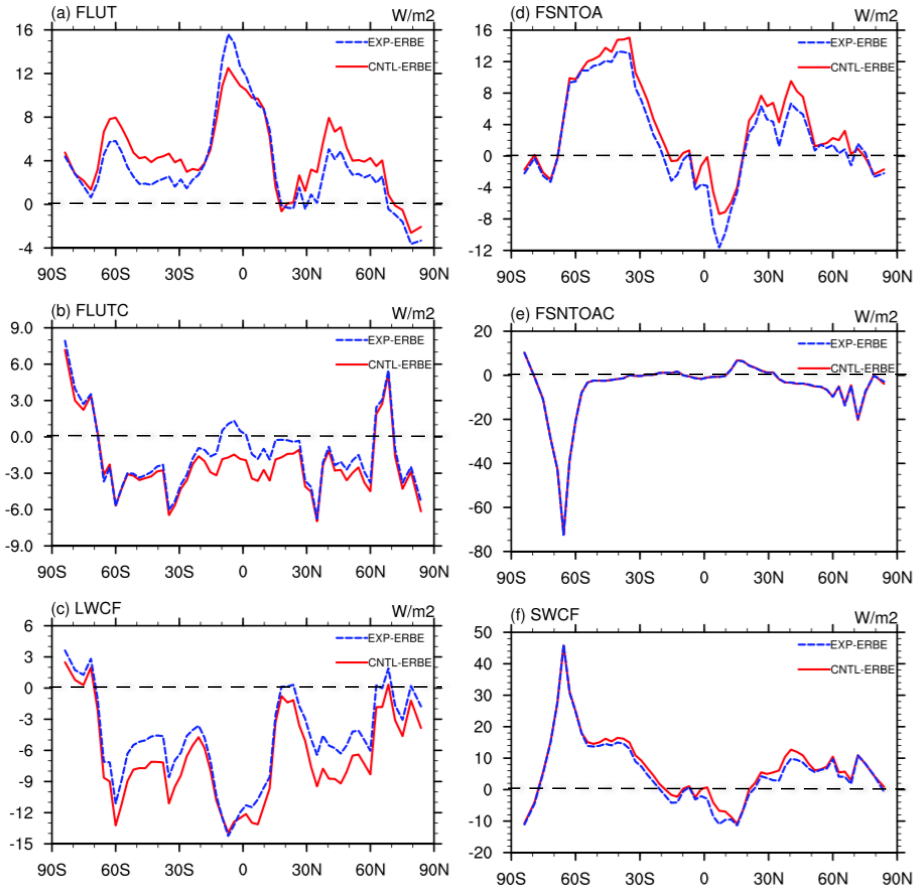


Figure 7. Meridional distributions of the annual mean difference between EXP/CNTL and observations of TOA outgoing longwave radiation **(a)**, TOA clearsky outgoing longwave radiation **(b)**, TOA longwave cloud forcing **(c)**, TOA net shortwave flux **(d)**, TOA clearsky net shortwave flux **(e)**, and TOA shortwave cloud forcing **(f)**.