RESPONSE TO REVIEWER #2 FOR GEOSCIENTIFIC MODEL DEVELOPMENT: MANUSCRIPT EGUSPHERE-2024-3770

We thank Reviewer #2 for the thoughtful and constructive feedback. This response document provides a response to each specific comment. Reviewer comments are in *blue italics*, author responses are in black, and changes to the manuscript are marked in red with line numbers referring to those in the revised manuscript.

Reviewer #2

This study presents a derivative-free optimization framework for tuning climate model parameters. The framework was applied to the GAMIL3 atmospheric model and evaluated for both 10-parameter and 20-parameter cases. The study assessed the framework's effectiveness in terms of the initial selection of model parameter values and found that the initial selection of model parameter values considerably affects the tuning results. The study also evaluated the effectiveness of applying the optimized model parameters, derived from the atmospheric model, to an atmosphere-ocean coupled climate model. Model parameterization optimization and model tuning are important aspects in the climate modeling community. The paper is well written and worth publishing. However, to benefit a wider modeling community, some issues need to be addressed and further clarification is necessary.

Reply: We thank the reviewer for their helpful and constructive comments, and have revised the paper accordingly.

Comment 1: *L174-175: Please provide more details about the initial trust region and parameter constraints. Is there any difference between parameter constraints and parameters' plausible ranges?*

Reply: We have revised the wording related to “parameter constraints” to clarify that it refers to constraints applied to the simulated variables, which is distinct from the physical parameters we tuned in this work. We have added the following explanation to the manuscript:” In the initialization of DFO-LS, we use the default parameter settings provided by the DFOLS software package, including the specification of the initial trust region, which is an algorithm parameter that governs the size of the local search area. Any constraints on the simulated variables are also specified at this stage. The initial trust region radius (*rhobeg*) is set to 0.18 (normalized to parameter ranges) based on sensitivity tests. This choice ensures that the first iterations explore locally without overstepping physical plausibility, balancing efficient convergence and sufficient sampling of the parameter space (Cartis et al., 2019). In addition, we apply a constraint to a simulated variable using a parameter μ, which determines the weighting of the constraint term (1/(2μ); see Supplementary S1). In this study, following Tett et al (2017,

2022), this constraint is applied to the global average TOA netflux. To tightly constrain this variable, μ is set to 0.18 which corresponds to a total uncertainty of 0.15 W/m² somewhat higher than the observational error of 0.1 W/m².” (Lines 177-189). We have also added further clarification regarding the distinction between the constraints to the simulated variables and plausible parameter ranges, as follows:” While the plausible ranges are defined as the maximum physically meaningful bounds (e.g., *rhcrit*: 0.65–0.95), the constraint on the global average TOA net flux ensures it closely matches the observations after tuning.”. (Lines 292-294).

Comment 2: *L180: In each iteration of the optimization process, how many simulations are conducted?*

Reply: Thank you for the comment. We have added further clarification as follows:” In addition to the initial *K*+1 simulation runs required to initialize the DFOLS algorithm for a *K*-parameter case, each iteration typically involves 1-3 additional model simulations, depending on the trust-region management strategy and the progress of the algorithm. The algorithm normally performs one simulation per iteration to evaluate a new candidate parameter set, but may conduct 3 simulations when the local quadratic model requires improvement or when the actual-to-predicted improvement ratio falls below zero (Cartis et al., 2019). Total evaluations include the initial runs plus all subsequent iterations evaluations.” (Lines 199-206).

Comment 3: *L215: A 30-year simulation is insufficient to fully evaluate the effectiveness of the modified model parameters in a fully coupled model.*

Reply: While we acknowledge that multi-century integrations would provide additional insights into the climate equilibrium state, our primary objective was to validate the transferability of AMIP-tuned parameters to a coupled framework, and a 30-year piControl simulation here in this study is scientifically sufficient to evaluate the effectiveness of the tuned parameters. We have added a discussion regarding this issue:” While we acknowledge that multi-century integrations would provide additional insight into the model’s equilibrium climate response, our primary goal was to test whether AMIP-tuned parameters remain valid in a coupled setup. For this purpose, a 30-year piControl run is scientifically adequate.

The results show that the model quickly reaches energy balance stability for both the 10- and 20-parameter cases (TOA net flux drift < 0.05 W m^{-2} per decade) and that ocean heat content drift remains minimal (< 0.008 °C per decade) after year 15, indicating that the system achieves a quasi-equilibrium state. This timescale is reasonable, since the upper ocean—where much of the adjustment occurs—has a relatively short adjustment timescale of about 1–5 years. The stabilized climate indicators and consistent system behavior (Figs. 9 and 10) confirm that the tuned parameters yield a credible coupled climate without introducing systematic drifts. Similar integration lengths have been used in other studies (e.g., Tett et al., 2017). While longer runs could refine the equilibrium further, they are unlikely to change our main conclusion that the parameter transfer is robust. ” (Lines 715-729).

Comment 4: *L226-228: \theta is not defined.*

Reply: Revised the text to:” we separate the analysis into four regions based on latitude (θ, defined as positive northward from the equator)” (Lines 267-268).

Comment 5: *L230-231: _TROPICALLAND, _TROPICALOCEAN, _NHX and _SHX are not defined*

Reply: Revised the text to:” While most variables are divided into four regions—labeled _TROPICSLAND (tropical land: 30° S–30° N over land), _TROPICSOCEAN (tropical ocean: 30° S–30° N over ocean), _NHX (Northern Hemispheric extra-tropics: >30° N), and _SHX (Southern Hemispheric extra-tropics: <–30° S)—each with its own target and uncertainty.” (Lines 272-276).

Comment 6: *L236: LAT is not defined*

Reply: Revised the text to:” Land Air Temperature (LAT)” (Line 281).

Comment 7: *L237-238: Please clarify how the uncertainty is derived from the absolute error*

Reply: Thank you for the reminder. In Section 2.4, we have clarified the different data sources used for each variable. To further improve clarity regarding our methodology, we added the following explanation:” The second matrix estimates the uncertainty of observations (C_0), which set to be diagonal, assuming no correlation between different

observations, and its values are derived from absolute difference between the two available datasets for each variable after regridding and area-weighting … …For the four radiation variables (OLR, OLRC, RSR, and RSRC), uncertainties are based on the estimates from Loeb et al. (2018)." (Lines 316-327).

Comment 8: *L250: I can't find them in the last column of Table 2*

Reply: Revised the text to "the first column".

Comment 9: *L405-407: The tuning process of the 20-parameter case was affected by using the same initial perturbations for the original 10 parameters. It is important to evaluate the effectiveness of the tuning method in terms of adding more parameters by comparing the 10-parameter and 20-parameter cases with independent initial parameter perturbations*

Reply: In our original experimental design, we intentionally maintained identical initial perturbations for the first 10 parameters in both the 10- and 20-parameter cases to establish a controlled comparison of how expanding the parameter space affects optimization outcomes. By holding the initial perturbations constant for these shared parameters, we ensured that any differences in the final tuned results could be directly attributed to the inclusion of additional parameters rather than variations in initialization.

However, in direct response to the reviewer's comment, we conducted a new experiment with completely independent initial perturbations for the 20-parameter case as a complementary. Since the optimized parameters from this experiment show quite similar performance to the original 20-parameter case, we have added this results to the discussion and supplementary: "to assess how the number of tuning parameters affects the optimization process, we used the same initial perturbation runs for the ten shared parameters in both the 10- and 20-parameter cases, enabling a consistent evaluation of their sensitivity to the simulated results. While this approach allows a straight forward comparison, it may also constrain the optimization in the 20-parameter case by introducing bias into the initial search space. To address this potential limitation, we conducted additional experiments in which all twenty parameters were initialized with independent perturbations (Fig. S4–S6) by adjusting the *rhobeg* parameter in the DFO-LS algorithm from

its default value of 0.18 to 0.23. These additional experiments yielded several important insights that strengthen our original conclusions. First, although the optimized parameter values in the new 20-parameter case differ somewhat from those in the original setup, most shift in the same direction relative to the default values (Fig. S4). Moreover, the optimization consistently converged to similar cost function values (2.68 vs. 2.87), despite differences in the initial perturbations and optimization pathways, highlighting the robustness of our tuning framework. Second, both approaches produced nearly identical simulation performance in the 10-year AMIP and 30-year piControl experiments (Fig. S5–S6), despite relying on different parameter sets. This suggests that the performance in the 20-parameter case may be dominated by a subset of the most sensitive parameters, such as *Dcs, rhcrit, c0_conv,* and *cmftau*, which have been shown to strongly influence the simulated results. These findings provide strong evidence that our conclusions regarding the robustness of the optimization and the effect of increasing the number of tuning parameters remain valid." (Lines 837-858).

Comment 10: *L416-417: What does "the initial 20 runs" refer to? Are these the initial perturbation runs conducted before the optimizing iterations begin? If so, please clarify this point. It appears that both the 10-parameter and 20-parameter cases achieve nearly the same STABLE performance by the 21 iterations. Does this mean the total number of runs for the two cases are 31 and 41 runs, respectively?*

Reply: The reviewer has raised an important point that warrants further clarification. Indeed, the initial 11/21 runs mentioned in the text refer to the perturbation runs conducted prior to the start of the optimization iterations. We have added the clarification to the Methods section; please refer to Comment 7 in our response to Reviewer #1.

Regarding the second comment—"*Does this mean the total number of runs for the two cases are 31 and 41 runs, respectively?*" —yes, the total number of model evaluations includes both the initial perturbation runs and the subsequent optimization iterations. For the two cases shown in Fig. 3, a total of 35 simulations (11 initial + 24 iterations) were conducted for the 10-parameter case, and 41 simulations (21 initial + 20 iterations) for the 20-parameter case. We have clarified this more explicitly in the revised manuscript by

focusing on the total number of iterations required to reach the minimum cost function value:" In the 10-parameter case, the optimization required 29 total model evaluations (11 initial perturbation runs + 18 iteration runs), reaching the lowest cost function value of approximately 3.5" (Lines 408-410) and "The system required a total of 31 runs (21 initial perturbation runs + 10 iteration runs) to reach the lowest cost function value (2.87), which is only two more than that required for the 10-parameter case." (Lines 471-473)

Comment 11: *L448: In an AMIP simulation, sea surface temperatures are specified, so ENSO (El Niño-Southern Oscillation) is not a suitable example in this context*

Reply: Thanks for pointing this out. We have revised the sentence to:" Although our cost function explicitly accounts for internal variability (Eq. 1), tuning and evaluating the model using only a one-year simulation may still introduce uncertainties due to atmospheric internal variability (Bonnet et al., 2025), such as phase shifts in the North Atlantic Oscillation (NAO) or stochastic tropical convection patterns like the Madden-Julian Oscillation." (Lines 502-506)

Comment 12: *L456-461: Does this indicate that the tuned results are tied to a specific climate background*

Reply: We acknowledge the reviewer's point regarding the tuning results for some variables, such as MSLP, which are somewhat tied to the specific climate background of the tuning period. However, most other variables (e.g., T500, RSR, NETFLUX) showed consistent improvements across both periods, demonstrating robustness against interannual variability. We have added further discussion on this in the manuscript and suggested that future work could explore tuning based on multi-year composites to better assess the generalizability of the results:" This temporal inconsistency suggests that certain parameter adjustments may be sensitive to the specific climate state of 2011, which was characterized by a moderate La Niña. In contrast, variables such as T500, RSR, and NETFLUX exhibit consistent improvements across both simulations, indicating a robust response to parameter tuning that is less dependent on interannual variability " (Lines 519-523) and added some discussion; please refer to Comment 16 in our response to Reviewer #1.

Comment 13: *L466-467: replace "equilibrium" with "energy balance"*

Reply: Replaced.

Comment 14: *L471: Why are MSL, RSRC, and LRC difficult to tune?*

Reply: We appreciate this technical question. The challenges in tuning MSLP and the two clear-sky radiation variables primarily stem from the gravity wave drag parameterization and the greenhouse gas effect related to water vapor. We have added a detailed explanation of these issues in the revised manuscript:" Specifically, MSLP is highly sensitive to unresolved gravity wave drag processes (Sandu et al., 2015; Williams et al., 2020), which were not included in our parameter tuning. Previous experiments with the IFS model indicate that increasing orographic and surface drag in the Northern Hemisphere can reduce MSLP biases (Kanehama et al., 2022). While the global mean OLRC is similar across cases due to regional compensation (Fig. 5d), the meridional distribution reveals notable differences (Fig. 7d). In the tropics, increased upper tropospheric water vapor—particularly in the 20-parameter case (Fig. 9a–9b)—enhances the greenhouse effect and reduces outgoing clear sky longwave radiation. In contrast, decreased water vapor in high-latitude regions, especially in the 20-parameter case, leads to increased OLRC. RSRC remains nearly unchanged across all simulations due to the use of identical surface albedo." (Lines 560-571)

Comment 15: *L474: OSRC is not defined*

Reply: Revised to RSRC.

Comment 16: *L476: TEMP@500 has been profoundly affected by tuning. Please explain the physical causes*

Reply: We thank the reviewer for highlighting this important point, which was previously underemphasized in the manuscript. We have revised the text accordingly. As shown in Fig. 8, nearly all of the first 10 parameters have a significant impact on TEMP@500, with adjustments to *rhcrit* and *Dcs* exerting the greatest influence in the 10- and 20-parameter cases, respectively. In this paper, we illustrate their potential impact from two perspectives: (a) convective heating profiles and (b) the radiative effects of upper-

tropospheric ice clouds—both of which are key drivers of the mid-tropospheric thermal structure. Of course, we acknowledge that different parameters may influence the simulated variables through different pathways, and while exploring these effects would be valuable, it lies beyond the scope of this study.

The physical explanations are presented in the manuscript for the 10-parameter case:" Low-level clouds strongly reflect shortwave radiation, producing a cooling effect. Therefore, a reduction in low-level clouds allows more shortwave radiation to penetrate the lower atmosphere, reducing outgoing shortwave radiation to space (blue lines in Fig. 5e and 7e) and warming the region (blue lines in Fig. 5a and 7a; Fig. 9e), including near the surface (blue lines in Fig. 5g)." (Lines 609-613), and for the 20-parameter case:" Specifically, clouds with higher ice content trap more OLR from the Earth's surface, potentially amplifying the greenhouse effect by retaining more infrared radiation (red lines in Fig. 6c and 8c). This results in a warming effect, particularly at lower atmospheric levels and even near the surface, especially during nighttime or in polar regions (red lines in Fig. 5a, 5g, 7a, and 7g; Fig. 9f)." (Lines 619-624).

Comment 17: *L479-480: Please add some discussion on how to tune the model performance for OLR and PRECIP*

Reply: Thank you for pointing this out. There was an incorrect expression in the original manuscript. While both optimized cases show worse PRECIP performance compared to the default case—particularly the 20-parameter case—the OLR for the 10-parameter case remains quite close to that of the default model. We have revised the original sentence to better emphasize the OLR and PRECIP performance differences, especially in the 20-parameter case:" In the 20-parameter case … …Both OLR and Lprecip perform notably worse than in the default case, with both variables being too low compared to the observations." (Lines 534-538). Additionally, we have included a discussion on possible tuning methods for these variables:" *ccrit, which sets the minimum turbulent threshold for triggering shallow convection, affects both OLR and Lprecip in a manner similar to Dcs* ……. Increasing ccrit suppresses shallow convection by requiring stronger turbulence to initiate cloud formation, thereby reducing low-level cloud cover. This reduction enhances outgoing longwave

radiation and surface solar heating, which in turn promotes evaporation and increases Lprecip. Therefore, adjusting Dcs and ccrit in future work may offer a promising approach for improving the simulation of OLR and Lprecip, both of which are underestimated relative to the default case." (Lines 617-636).

Comment 18: *L534-542: The 10-parameter case shows a larger difference in TOA outgoing shortwave flux (RSR) compared to the 20-parameter case relative to the default case (Fig. 4e and 6e). However, the 20-parameter case exhibits a larger difference in cloud compared to the 10-parameter case relative to the default case (Fig. 8d-e). Please explain this discrepancy*

Reply: We thank the reviewer for identifying this behavior, which we agree should have been stated more explicitly. The apparent discrepancy between changes in RSR and cloud fraction arises from competing microphysical and radiative effects in the 20-parameter case. We have added a detailed explanation for this in the revised manuscript:" Additionally, raising the autoconversion threshold from ice to snow is expected to allow more ice to remain in the atmosphere, directly leading to a reduction in precipitation (red line in Fig. 5h), and increased cloud optical thickness, thereby enhancing the reflection of incoming shortwave radiation. This enhanced reflectivity partially offsets the impact of reduced low-level cloud cover on the RSR in the 20-parameter case, leading to a smaller decrease in RSR compared to the 10-parameter case (Fig. 5e and 7e), consistent with known radiative differences among cloud types (Chen et al., 2000)." (Lines 624-631).

Comment 19: *L594-613: anomalies => biases*

Reply: Revised.

Comment 20: *L565-619: Does the coupled model directly utilize the optimized parameters from the AMIP simulations? If so, the TOA energy imbalance caused by the optimized parameters would eventually lead to climate drift in the long-term integration of the coupled model. This undermines the rationale and effectiveness of applying parameters tuned for an atmospheric model to an atmosphere-ocean coupled model. Meanwhile, a 2 W/m² energy imbalance at TOA is not a "slight energy imbalance" as stated in the abstract*

Reply: The parameter sets used in the coupled model were directly adopted from the AMIP-optimized results, which is a common practice in climate model tuning (Zhang et al. 2015; Hourdin et al., 2016; Tett et al., 2017;). The net flux at the TOA in AMIP simulations includes the effect of greenhouse gases, whereas this effect is not represented in the piControl (coupled) runs. We have incorporated this detailed clarification into the revised manuscript:" based on the assumption that parameters performing well under observed forcings (e.g., prescribed SST, sea ice, and greenhouse gases) in the standalone atmospheric model will also improve performance in the coupled system. In our case, the TOA energy imbalance in the AMIP run mainly results from the radiative forcing of greenhouse gases, which trap outgoing longwave radiation. Since the piControl experiment is forced by constant pre-industrial greenhouse gas levels, this radiative effect is absent. Therefore, if the AMIP-tuned parameters correctly capture this effect, the coupled model under piControl conditions should yield a near-zero TOA net flux, as expected." (Lines 241-249).

Regarding the relatively large energy imbalance at the TOA observed in the coupled runs for both optimized cases, we acknowledge this as an intrinsic limitation of the atmospheric model. This imbalance primarily originates from a persistent energy imbalance in the atmospheric component's dynamical core, which is carried over from the AMIP simulations into the piControl runs. We have included a detailed discussion of this issue in the revised manuscript. Please refer to our response to Reviewer #1, Comment 14, for further details.

In addition, we have revised the abstract to:" Additionally, evaluations of the coupled model with optimized parameters showed, compared to the default parameters settings, reduced climate drift, a more stable climate system, and more realistic sea surface temperatures, despite an overall energy imbalance of 2.0 W/m², approximately 1.4 W/m² of which originates from the intrinsic imbalance of the atmospheric component, and the presence of some regional biases." (Lines 33-38)

Comment 21: *L767: forecasts -> prediction*

Reply: Revised.