RESPONSE TO REVIEWER #1 FOR GEOSCIENTIFIC MODEL DEVELOPMENT: MANUSCRIPT EGUSPHERE-2024-3770

We thank Reviewer #1 for the helpful feedback. In our response, 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 #1

Overview: I am reviewing the revised version of this manuscript. While the work is well presented, I remain concerned about the robustness and interpretability of the proposed tuning framework.

Reply: We thank the reviewer for the comments and insights.

Comment 1: *The study relies on 1-year simulations to assess the impact of parameter perturbations. Is one year truly sufficient to obtain stable and representative results? Climate models often require longer integrations for meaningful statistics.*

Reply: While we agree that longer integrations are generally preferable for robust statistics, particularly for variables with high temporal variability (as discussed in the paper, Lines 829–847), our one-year window represents a practical compromise between computational cost and scientific rigor. Here, we outline the rationale for this choice and summarize multiple lines of evidence demonstrating that the tuning results remain robust and stable despite the short optimization period:

1. Central to our approach is the explicit incorporation of internal variability into the optimization framework. Using a 20-member perturbed-initial-condition ensemble, we estimated the model's internal variability covariance matrix (C_i) and combined it with the observational uncertainty (C_0) to obtain the total covariance matrix $C = C_0 + 2C_i$ used in the cost function. This formulation "ensures that simulated observations within the range of internal variability receive reduced penalties, guiding the optimization to correct systematic biases while avoiding overfitting to random climatic fluctuations". (Lines 325-327).

2. (Lines 829-847) **Beyond this methodological safeguard, we applied a rigorous multi-stage validation protocol to assess temporal representativeness**. When tested in an independent decade-long AMIP simulation (2005–2014), the parameters optimized from a single year retained their improved performance across most key variables, indicating that the results are not confined to the specific climate conditions of 2011. Furthermore, when transferred to a fully coupled configuration, these parameters yielded a more stable energy budget, reduced climate drift, and improved sea surface temperature patterns over a 30-

year pre-industrial control run. Consistent improvements across distinct model configurations and extended timescales provide strong evidence that the one-year tuning captures robust physical improvements rather than transient or case-specific artifacts.

Comment 2: *In this manuscript, the optimized parameters happen to lead to an overall improvement. But is such improvement guaranteed, or could it be a matter of chance? What if, in other cases, the parameters suggested by DFO-LS produce substantially worse agreement with observations? Should one then switch to a different optimization algorithm, or select alternative parameter values manually? I am concerned about this possibility, as it complicates attribution and reduces confidence in the robustness of the proposed method.*

Reply: We agree that success in any single tuning exercise cannot be guaranteed, as outcomes depend on structural errors, cost function design, and initial parameter values. However, generally speaking, we emphasize that the algorithm has a mathematically rigorous construction and is provably convergent to a (local) minima for a wide variety of smooth cost functions, and provably guaranteed to provide improvement upon the initial starting guess unless we are at a (local) solution. Furthermore, multiple aspects of our study indicate that the observed improvements are robust rather than fortuitous. Below, we address the reviewer's concerns point by point:

1. The success observed in this study is not attributable to chance. Model parameters were constrained within physically plausible ranges based on expert judgment, preventing unrealistic excursions in parameter space (Lines 246–249). Although the algorithm efficiently minimized the cost function from different initial conditions, sensitivity experiments showed that expert-informed initialization produced the most physically consistent results (Lines 790–793), illustrating that the framework complements rather than replaces physical insight. Most importantly, "validation was performed through extended independent decadal AMIP (AMIP2005-2014) simulations and 30-year coupled piControl simulations. Consistent performance across timescales and model configurations confirmed that the tuning corrected systematic biases rather than overfitting" (Lines 897–901).

2. We acknowledge that the possibility of suboptimal outcomes is an important practical consideration. Our framework, however, is not intended as a black box but as a

tool for systematic exploration. If DFO-LS suggests parameter sets that degrade performance, the solution is not a binary choice between switching algorithms and manual intervention, as alternative optimization methods do not necessarily guarantee better results. Instead, we "conducts a comprehensive diagnostic analysis—examining spatial patterns, process-level responses, parameter sensitivities, and multi-variable metrics—to assess the physical credibility of each solution" (Lines 210–212). When results are unsatisfactory, one can reinitialize the optimization from an alternative expert-informed starting point, adjust the cost-function weights to reinforce key constraints, or employ other algorithms (e.g., Gauss–Newton) previously tested in Tett et al. (2013). "This structured yet flexible workflow transforms the modeller's role from manual trial-and-error to managing and interpreting automated explorations, thereby improving both the traceability and objectivity of the modeling process" (Lines 212–215).

Comment 3: *The framework is demonstrated on one model version. During model development, new processes are added, and physics schemes evolve. In such cases, previously tuned parameters may need to be retuned, and the "optimal" set may not carry over. Does this framework really help us understand why the model is biased—e.g., due to misrepresented physical processes—or does it simply provide a temporary calibration that may not generalize across versions? This is an important issue for assessing the scientific value of the approach.*

Reply: The reviewer's comment highlights the long-term scientific value and diagnostic power of our tuning framework. We agree on two key points: first, that an optimal parameter set derived for one model version is unlikely to transfer directly to future versions with substantial structural changes; and second, that the ultimate goal of model development is to identify and correct misrepresented physical processes rather than merely mask biases through parameter adjustment. Nevertheless, we emphasize that our framework is more than a tool for temporary calibration—it actively facilitates physical understanding and model improvement for the following reasons (also clarified in the Discussion section of the paper):

1. **Beyond its role in model calibration, the framework provides rich diagnostic information throughout the optimization process**. While the output is a set of parameters, the more valuable contribution lies in the diagnostic insights generated during tuning. The framework produces information critical for process-oriented model evaluation, including parameter sensitivities and compensating errors among model parameterizations. For example, the Jacobian matrix (Fig. 8) shows how each model variable responds to specific parameters. When a bias can be corrected by adjusting parameters from multiple physically distinct schemes, this indicates compensating errors, a key step in diagnosing structural weaknesses. Sections 3.3 and 3.4 demonstrate how interactions between deep convection (*rhcrit*) and microphysics (*Dcs*) parameters affect clouds, radiation, and precipitation, revealing the processes behind energy and hydrological biases. Furthermore, the fact that different parameter combinations can produce similarly acceptable results highlights a critical insight for assessing the robustness of climate projections.

2. **The calibration framework can also accelerate the model development cycle**. While the reviewer correctly notes that new physics schemes require re-tuning, this is precisely where our framework demonstrates its practical value. Traditional manual tuning is slow, labor-intensive, and often subjective, whereas our automated approach allows developers to efficiently assess and compare model versions. First, when a new physical scheme is introduced, the framework can rapidly generate an objectively tuned parameter set for the updated configuration, enabling systematic comparison with the previous version. Second, by analyzing optimized parameters and their sensitivities across versions, developers can identify the impacts of structural changes: if a new scheme requires substantially different parameter values or alters sensitivity patterns, this provides concrete evidence of how the modification affects model physics and dynamics. In this way, the framework serves as a magnifying glass for the consequences of model development.

We have also incorporated clarifications on this issue in the Discussion section to help readers better understand the significance of the tuning framework:" the proposed DFO-LS-based tuning framework presents a robust and efficient approach for enhancing climate model performance. By combining Jacobian estimation with sensitivity analysis, the framework

quantitatively maps how parameters affect key variables and thereby exposes compensating errors between physical schemes (for example, interactions between deep convection and microphysics). These parameter–variable mappings yield direct insight into model structural uncertainties and supply objective diagnostics that guide development. When model physics are changed, the framework supports rapid retuning and systematic inter-version comparison: systematic shifts in optimal parameter values then serve as concrete evidence of how structural modifications alter model behaviour. Implemented and exercised primarily by a single researcher within 12 months, the approach also demonstrates high human-resource efficiency and practical scalability. Although no single parameter set is expected to transfer unchanged across model generations, automating the exploration process transforms development from manual trial-and-error into an efficient, reproducible, and more objective workflow. Applied across GCMs, this methodology can accelerate model development, reduce parametric uncertainty, and improve the reliability of climate projections." (Lines 932-948).

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

The manuscript requires more minor clarifications.

Reply: We thank the reviewer for the detailed comments and have revised the manuscript accordingly.

Comment 1: *Lines 25-31: this part is obscure. Please rephrase it.*

Reply: We have rephrased it to "To evaluate its performance, two main tuning experiments were conducted, targeting 10 and 20 parameters, respectively. In addition, three sensitivity experiments tested the effect of varying initial parameter values in the 10-parameter case. Both tuning experiments achieved a rapid reduction in the cost function. The 10-parameter optimization improved model accuracy for 24 of 34 key variables, while expanding to 20 parameters yielded improvement for 25 variables, though some structural model biases appeared" (Lines 26-32).

Comment 2: *Lines 74-75: please rephrase "the accuracy and skill of climate model outputs".*

Reply: Rephrased to "Appropriate parameter tuning enhances the accuracy and skill of climate models by optimizing parameter values to better match observations or high-resolution simulations used as calibration targets" (Lines 75-77).

Comment 3: *Line 92: "7 and 14 parameters were estimated", what kind of parameters?*

Reply: Revised to "7 and 14 parameters related to the convection, cloud microphysics, and boundary-layer dynamics (Yamazaki et al., 2013) were estimated using variants of the Gauss-Newton algorithm (Tett et al., 2013) to minimize the differences between simulated and observed large-scale, multi-year averaged net radiative fluxes" (Lines 93-97).

Comment 4: *Line 96: "focusing on seven parameters", what kind of parameters?*

Reply: Revised the text to:" Zhang et al. (2015b) employed an improved downhill simplex method to optimize seven parameters selected from the convection and cloud-fraction parameterization scheme, and reported successful improvement of an atmospheric model's performance" (Lines 98-100).

Comment 5: *Lines 145-148: these have been well introduced in Section 2.2.*

Reply: Deleted.

Comment 6: *Lines 169-171: it is suggested to name each module directly in Figure 1 for better understanding. "Python 3.8+" might not necessarily be displayed on the framework.*

Reply: Please see the revised picture shown below.

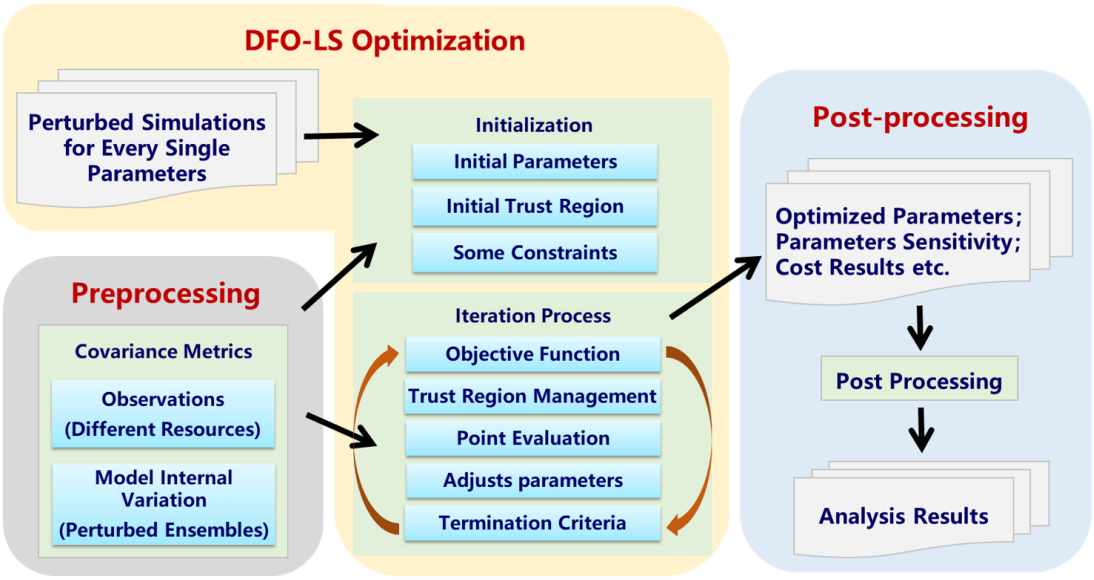

Figure 1. Automatic tuning framework structure. Perturbed simulation results for each parameter are used for sensitivity analysis and determining the trust region size. Two key covariance metrics—observational error and model internal variation—help adjust parameter values in the objective function. The DFO-LS algorithm optimizes the parameters, and the post-processing module analyzes sensitivity, cost function results, and generates visualizations.

Comment 7: *Line 187: netflux => net flux.*

Reply: Corrected.

Comment 8: *Lines 243-249: this section comes across as jarring in the context of the surrounding text.*

Reply: Revised the text to "A 30-year piControl simulation (Eyring et al., 2016) was then conducted to assess the model's long-term energy balance and stability under constant pre-industrial forcings. This experiment tests whether parameters performing well under observed forcings in AMIP simulations—such as prescribed SSTs, sea ice, and greenhouse gases—can also improve coupled performance. In AMIP runs, the TOA energy imbalance

mainly results from greenhouse gases forcing, which traps outgoing longwave radiation. Under piControl conditions, where pre-industrial greenhouse gas concentrations are fixed, this radiative effect is absent; thus, if the AMIP-tuned parameters are physically consistent, the coupled model should yield a near-zero TOA net flux" (Lines 298-306).

Comment 9: *Line 257: it is not clear what is "the first 10 parameters".*

Reply: We have reordered Sections 2.2 and 2.3 so that the tuning parameters are introduced first, followed by the experiment descriptions, allowing the "first 10 parameters" mentioned in Section 2.2 to be clearly defined.

Comment 10: *Line 260: It is suggested that the experiment names (AMIP2011, AMIP2005-2014, and piControl) be included in Figure 2. Additionally, clarification is needed regarding the differences between "1-year optimization for 10 parameters" and "1-year optimization by varying 10 parameters," as all parameters are varied during the optimization process.*

Reply: We thank the reviewer for the suggestions and have revised Figure 2 and its caption accordingly.

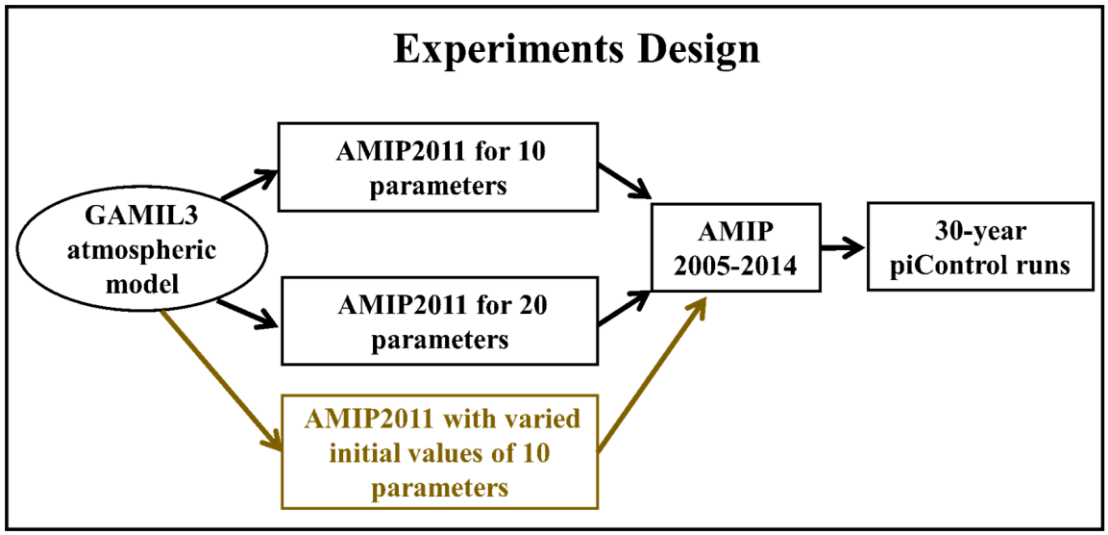

Figure 2. All experiments conducted in this study, including the AMIP2011 optimization runs for 10- and 20-parameter cases, the AMIP2005-2014 simulations using the optimized parameter sets, and the 30-year piControl simulations. Note that the piControl simulations were not performed for the sensitivity experiments that varied the initial values of the 10 parameters (shown in brown).

Comment 11: *Line 272: what does the suffix "_DGM" stand for in Table 1?*

Reply: We thank the reviewer for pointing it out. The suffix "_DGM" stands for "delta global mean". In response, we have added the following clarification in the manuscript "For the MSLP variable, regional mean values are expressed as anomalies relative to the global mean (delta global mean, denoted by the suffix "_DGM"), obtained by subtracting the global average from each regional mean". (Lines 232-235).

Comment 12: *Lines 408-410: It is recommended that the x-axis of Figure 3b & 3c start at 1 instead of 0 to better reflect the number of runs. Additionally, while the x-axis is labeled "Iterations," it actually represents the total number of runs, including both initial perturbations and iteration runs.*

Reply: The starting value of the x-axis and the axis title have been revised as suggested.

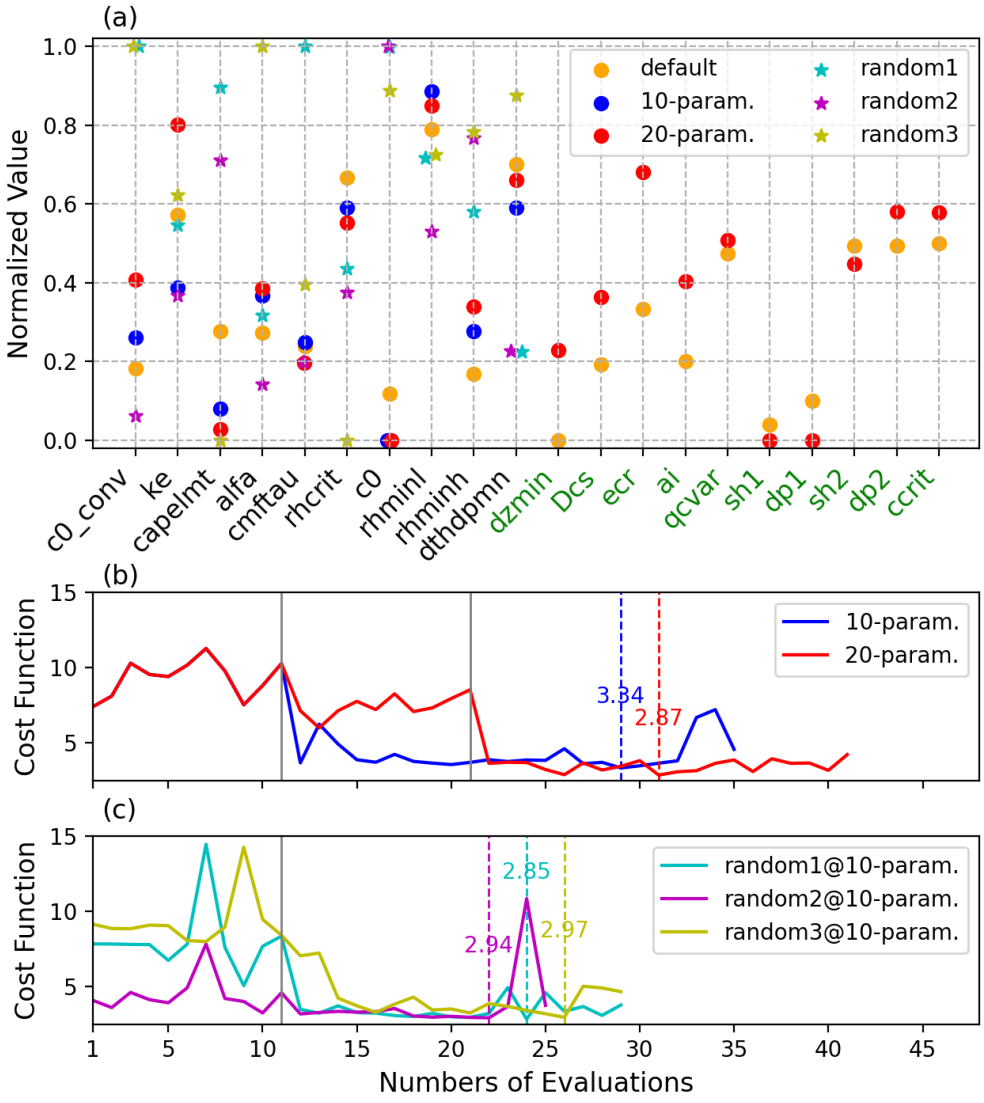

Figure 3. Normalized values of tuning parameters for the default and all five optimized cases (a);

changes in the cost function values over numbers of evaluations for the two main optimized cases (b) and the three sensitivity experiment cases (c). The vertical solid lines indicate the 11 and 21 runs from the initial perturbation phase, while vertical dashed lines mark the iterations at which the cost function reach its minimum.

Comment 13: *Lines 410-411: does the cost function drop rapidly from about 7.5 to 3.5 in the first iteration run? If so, please revise "during the initial perturbation phase" to "in the first iteration run."*

Reply: Yes, the reviewer is correct; it has been revised accordingly.

Comment 14: *Lines 433-434: it is recommended to delete "patterns".*

Reply: Deleted.

Comment 15: *Line 450: please rephrase "patterns".*

Reply: Revised to "while most variables exhibit similar behaviors to those of the Optimized set, notable differences are observed in T2M and Lprecip." (Lines 461-462).

Comment 16: *Line 470: "after the initial 20 perturbation runs", as in the 10-paramter case, does the cost function drop rapidly in the first iteration run? If so, please clarify.*

Reply: Yes, the reviewer is correct; it has been revised accordingly "it is evident that the cost function dropped rapidly to a value very close to the minimum in the first iteration run, similar to the 10-parameter case." (Lines 480-482).

Comment 17: *Line 477: "the initial phase" => "the initial perturbation phase".*

Reply: Corrected.

Comment 18: *Line 515: "MSLP_TROPICSOCEAN_DGM improved by over 20", what does the suffix "_DGM" stand for? What's the unit for the value "20"?*

Reply: As noted in our response to Comment 11, the suffix "_DGM" denotes "delta global mean." Since the y-axis is calculated using the formula shown below, it represents a dimensionless value corresponding to the difference between the optimized and default simulations normalized by the standard error. We have revised the sentence in the manuscript for clarity as follows " while the MSLP_TROPICSOCEAN_DGM shows an improvement of more than 20 standard errors relative to observations in the 2011 simulation with the 10-parameter case, it deviates from the observations by over 10 standard errors in the 10-year simulation." (Lines 524-527).

$$Z = \frac{|V_{\text{Default}} - V_{\text{Observation}}| - |V_{\text{Optimized}} - V_{\text{Observation}}|}{Standard\ error}$$

Comment 19: *Lines 643-689: It's intriguing that the tuning procedures reveal an energy leakage in the atmospheric model. Although the tuning can't eliminate this issue, the fact that it manifests as a TOA energy imbalance in the piControl run is remarkable. Given that half of the tuning targets are TOA energy components, as outlined in Table 1, and considering that NETFLUX has been effectively tuned to the target value of 0.98 W/m² in the AMIP run, it remains curious how this energy leakage manifests in the AMIP run post-tuning and why it emerges as a TOA imbalance in the piControl run.*

Reply: The reviewer has accurately identified a key paradox that lies at the heart of the model's energy budget behavior. This can be explained by the difference in how the AMIP and coupled configurations handle energy fluxes.

In the AMIP configuration, the use of prescribed observed SSTs effectively turns the ocean surface into an infinite energy source or sink. As a result, any energy leakage originating from the atmospheric dynamical core is compensated by an implicit, non-physical energy flux through the ocean surface. This allows the TOA energy balance to be tuned to match the target (~0.98 W/m²), as the prescribed SSTs mask the model's internal energy conservation issue.

In the coupled simulation, however, this artificial compensation mechanism is removed. The ocean now dynamically participates in the energy budget, and the previously hidden atmospheric energy leakage can no longer be absorbed at the boundary. It is noteworthy that despite this ~2.0 W/m² persistent TOA imbalance, the coupled system reaches a quasi-stable state with minimal drift in surface temperatures—indicating that the energy leakage remains constant and the model maintains a new, albeit biased, equilibrium.

The reviewer's intuition is indeed supported by further evidence. Our ongoing work (not shown in this manuscript) actually focuses on applying the energy conservation correction proposed by Williamson et al. (2015) to the atmospheric dynamical core and repeating the tuning process. With this correction, the optimized parameters now yield a TOA energy imbalance much closer to zero (approximately 0.5 W m^{-2}) in the coupled model. This

confirms that the original ~2.0 W/m² imbalance indeed originated from intrinsic energy leakage in the atmospheric component—a bias that is masked in AMIP but exposed in a coupled framework.

In response to the reviewer's comment, we have clarified in the manuscript "Additionally, the optimized cases show a relatively large TOA energy imbalance (~2.0 W/m²) despite a well-tuned NETFLUX in AMIP runs, which originates from energy non-conservation in the atmospheric model's dynamical core. In the AMIP configuration, prescribed SSTs act as an infinite energy source/sink, masking this internal leakage in the dynamical processes. By contrast, the coupled system exposes the dynamical core's non-conservation as a stable but imbalanced energy state. This interpretation is supported by our ongoing experiments (not shown) following Williamson et al. (2015), where correcting energy conservation in the dynamical core reduced the TOA imbalance in the piControl runs to about 0.5 W m^{-2} within the same tuning framework. These results underscore that while parameter tuning can improve model fidelity, structural errors in the dynamical core—particularly its energy non-conservation—must be addressed to achieve physically consistent climate simulations."(Lines 876-887).