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

We thank Reviewer #1 for the thoughtful and constructive feedback. In this 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

This well-structured manuscript presents a novel approach for climate model-tuning and the results that such tuning yields for a given model (GAMIL3) under 3 different model configurations: 1 year AMIP for tuning , 10 year AMIP and 30 year coupled pre-industrial Control. The presented tuning method is potentially relevant for other climate models. The authors show that the DFO-LS method is able to systematically improve the 'a priori' model parameter values and that the improvements hold across the different model configurations. The text is well written, with some potential however for more precise and less verbose language. In general, the manuscript could improve by adding some comparison or references to similar past efforts on model tuning, but I acknowledge that often findings and results are quite model-specific.

Reply: We thank the reviewer for this comment.

Comment 1: *L45-46 Some references would be welcome.*

Reply: In the revised version, we have incorporated several relevant references to support this point:" In recent decades, significant progress has been made in advancing the major components of the Earth system—such as the atmosphere, ocean, land, and human systems (Prinn 2012; Bogenschutz et al., 2018; Fox-Kemper et al., 2019; Blockley et al., 2020; Blyth et al., 2021)—as well as in developing the coupling techniques required to form fully integrated ESMs (Valcke et al., 2012; Smith et al., 2021; Liu et al., 2023)." (Lines 46-51).

Comment 2: *L186: Not strictly necessary, but perhaps having a sketch showing the sequence of experiments performed would help the reader.*

Reply: We have added a flow chart in the revised manuscript, as shown below.

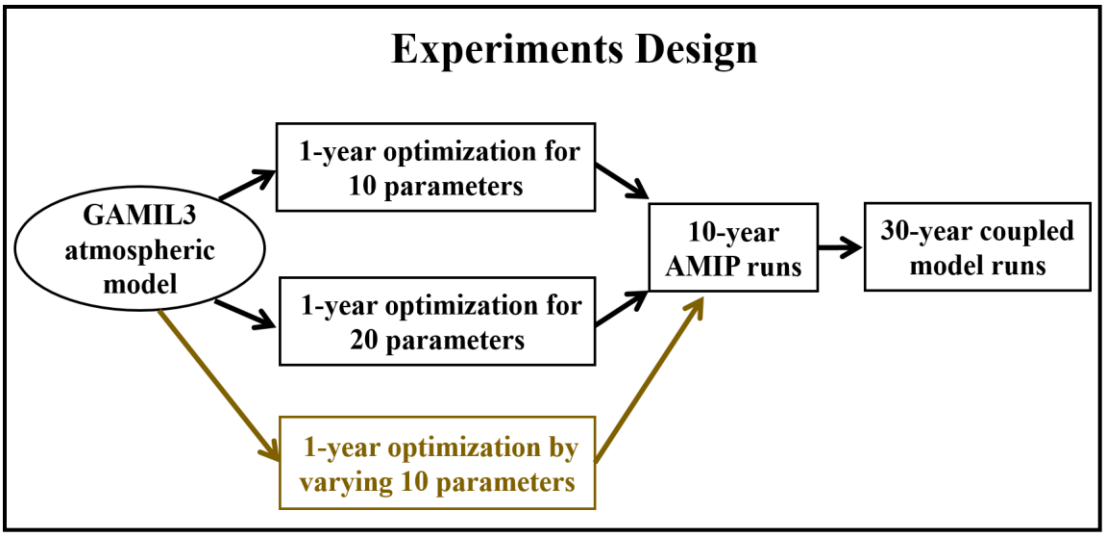

Figure 2. Overview of all experiments conducted in this study, including the 1-year AMIP (AMIP2011) optimization runs for the 10- and 20-parameter cases, the 10-year AMIP (AMIP2005-2014) simulations, and the 30-year piControl simulations using the optimized parameter sets. Note that piControl simulations were not conducted for the varying 10-parameter cases, which are indicated in brown.

Comment 3: *L186: The text has no literature reference for GAMIL3. If no documentation exists for this model version, a more detailed description of it would be needed, as an Appendix if needed. The current description between L187-203 is vague and full of ambiguities ('updates to the planetary boundary layer scheme', 'GAMIL3 integrates several parametrizations recommended by CMIP6').*

Reply: GAMIL3, with a 2-degree (180 × 80 grid) horizontal resolution, is the atmospheric component of FGOALS-g3; both have completed the CMIP6-related experiments (Li et al., 2020a, b). In this study, we use its higher-resolution 1-degree (360 × 180 grid) version, which is identical to GAMIL3 except for the time step of the dynamical core. Accordingly, we have removed the introduction of GAMIL2 and revised the relevant sections to place greater emphasis on GAMIL3: "In this study, we employ GAMIL3, which adopts a finite difference dynamical core and a weighted equal-area longitude-latitude grid to maintain numerical stability near the polars without the need for filtering or smoothing (Wang et al., 2004; Li et al., 2020a). GAMIL3, with an approximate 2° (180×80) horizontal resolution, serves as the atmospheric component of the Flexible Global Ocean–Atmosphere–Land System Model Grid-point Version 3 (FGOALS-g3), which participated in CMIP6 (Li et al., 2020b). For this

study, the model's horizontal resolution is refined to about 1° (360 × 160), with 26 vertical σ-layers extending to the model top at 2.19 hPa. To ensure numerical stability at the higher resolution, the dynamical core time step is reduced from 120s to 60s, while the physical parameterizations and their time step (600s) remain unchanged. As in many other climate models (e.g., Santos et al., 2021; Wan et al., 2021; Schneider et al., 2024), the performance of GAMIL3 is sensitive to the resolution, the model time step, and the coupling frequency between dynamics and physics. Therefore, it is necessary to re-tune the uncertain parameters for the new 1° configuration."(Lines 211-224).

Comment 4: *L280: Is there any reason or reference why you would give twice as much weight to C_i than to C_0?*

Reply: We applied a doubling factor to the variability component because both model simulations and observations contain internal variability. Assuming two independent sources of variability justifies using twice the estimate from control simulations. This reflects a conservative assumption that both sources contribute comparable levels of noise. This approach follows the practice of Tett et al. (2022), and we have included this clarification in the revised manuscript:" Consistent with Tett et al., (2022), we account for internal variability in both model simulations and observations by doubling the model-based estimate, reflecting a conservative assumption of comparable noise contributions." (Lines 331-333).

Comment 5: *L296: put the definition of the Jacobians in context. Why are you presenting it? where in the paper is used?*

Reply: We have added background information on the Jacobian prior to its introduction in the revised manuscript:" The Jacobian matrix, *J*, defined as the partial derivatives of the simulated outputs with respect to the parameters being optimized, is used to assess the influence of tuning parameters on the simulated variables." (Lines 348-350). The sensitivity of the tuning parameters to the simulated outputs is illustrated in Figs. 8 and 13, both of which analyze the parameters' impact on the modelled variables.

Comment 6: *L351: Why ke and captlmt are explicity mentioned? please explain*

Reply: Our intention here is to highlight the parameters that underwent substantial changes through tuning compared to others. We have revised the corresponding sentences to clarify this point: "In this experiment, several parameters—such as *ke* and *captlmt*—changed significantly from their default values, while *cmftau* and *c0* showed only small changes (Fig. 3a)" (Lines 403-405).

Comment 7: *L414: Any illustrative example of compensating errors in the model?*

Reply: We used the term 'compensating errors' to emphasize the underlying interactions whereby adjustments to one parameter can offset or amplify the effects of another. An example for *cmftau* is discussed in detail in the paper:" Although the 10-parameter case has a higher threshold for low level cloud formation than the 20-parameter case, Fig. 9c-9d shows the different result, which can be explained by the compensatory effects of other parameters. Optimized results indicate that *cmftau*, another key parameter, has a lower value in the 20-parameter case (~4284) compared to the default (~4800) and the 10-parameter case (~4931). This decrease in *cmftau* likely strengthens shallow convection while weakening deep convection, reducing upward water transport and RH throughout the troposphere, contributing to the decreased low-level cloud fraction (Xie et al., 2018) and further reducing precipitation (Fig. 5h). Consequently, the lower low-level cloud fraction in the 20-parameter case, compared to the 10-parameter case, reflects the compensatory effects of these key parameters, with the influence of the reduced *cmftau* outweighing that of *rhminl*." (Lines 598-609). For the parameter *Dcs*, its counteracting effects with the parameters *rhminl* are discussed in the paper:" 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). The *ke* parameter, has a contrasting effect on

OLR and RSR to the *capelmt* parameter, although its impact on most simulated variables is minor, as shown by the Jacobian (Fig. 8).

We have revised this sentence to be more precise:" These differences may be attributed to the compensating errors within in the model, where adjustments to one parameter can offset or amplify the effects of another—a phenomenon further explored in Section 3.3." (Lines 466-468).

Comment 8: *L503: I'd re-name this section as "Coupled model evaluation'*

Reply: We have revised the title of Section 3.4 to '**Coupled Model Evaluation**.' Since Section 3.3 focuses on providing a physical explanation of the tuning results for the 10- and 20-parameter cases, we have also updated its title to '**3.3 Impacts of Tuning on GAMIL3**' to better reflect its content and avoid potential misunderstandings.

Comment 9: *L521: lower rhcrit could, a priori, also enhance precip. Lower rhcrit would enhance convection and, this, precipitation. Even if it is not the case in the simulations, it may be worth being mentioned*

Reply: We agree with the reviewer that, in principle, a lower *rhcrit* should increase precipitation. However, our simulations show a net reduction, which is likely attributable to compensating effects such as moisture redistribution. We have added this discussion to the revised manuscript:" While a lower *rhcrit* threshold would theoretically enhance precipitation by promoting deeper convection, our simulations instead show an overall decrease in precipitation. This apparent discrepancy suggests the parameter's effect is modulated by compensating atmospheric processes. Specifically, enhanced vertical moisture transport (Fig. 9a-9b) reduces low-level humidity availability, thereby weakening updrafts and ultimately decreasing total precipitation (blue line in Fig. 5h)." (Lines 589-595).

Comment 10: *L533: contributing to the decreased low-level cloud fraction and further reducing precipitation (since this was mentioned in the previous paragraph)*

Reply: We have refined the sentence to better maintain the logical connection between cloud fraction and precipitation as follows:" contributing to the decreased low-level cloud fraction (Xie et al., 2018) and further reducing precipitation (Fig. 5h)." (Lines 605-606).

Comment 11: *L569: Describe for how long the coupled model was run, one can only infer it from the Figures*

Reply: Thank you for the reminder. We have added an explicit clarification in the manuscript:" To assess the impacts of atmospheric parameter tuning on coupled model performance, we conducted a 30-year piControl simulation using GAMIL3 coupled to land, ocean, and sea ice components (see Methods 2.2), analyzing the final 15-year period after model spin-up.". (Lines 639-642).

Comment 12: *L569: for coupled simulations it is quite relevant to explain how the land, and specially the ocean, were initialized. This is relevant because a perfect model should drift if the ocean is not correctly initialized, and you would not like to tune your model to compensate for an ocean-caused drift*

Reply: This is an important point. Apart from the difference in the resolution of the atmospheric component, we used the same model as FGOALS-g3, which participated in CMIP6. The initial conditions for the piControl run were derived from the climatological mean state of atmospheric reanalysis for the atmospheric model (default configuration), and from the equilibrated state of the OMIP simulation—a long ocean-only run forced by atmospheric reanalysis—for the ocean model. No prescribed initial conditions were used for the land component; instead, its state was generated during the coupled integration. To minimize the impact of potential initialization drift, the first 15 years were treated as a spin-up period and excluded from the analysis. This clarification has been added to the Methods section:" The initial condition for the atmospheric model was the climatological mean state from atmospheric reanalysis (default configuration), while the ocean model was initialized from the equilibrated state of an OMIP simulation (a long ocean-only run forced by atmospheric reanalysis). The land model was not provided with a prescribed initial condition; instead, its state was generated dynamically during the coupled integration. To minimize the influence of potential

initialization drift, the first 15 years were treated as a spin-up period and excluded from the analysis." (Lines 249-256).

Regarding the potential drift induced by the initial state, we have added the following discussion:" Drift may occur during the initial integration period due to inconsistencies between the OMIP-forced ocean state and the reanalysis-based atmospheric initial conditions. However, in both cases using atmosphere-optimized parameters, the system stabilized rapidly, and neither the TOA net flux nor ocean temperature exhibits significant trends beyond the initial adjustment period of a few years. A small long-term drift is still evident in Fig. 10d, which may be related to the adjustment of deep ocean processes. This demonstrates that the parameters optimized for the atmospheric model remain effective in the coupled system configuration, with no clear evidence of compensation for ocean-related drift." (Lines 681-689).

Comment 13: *L575: While the reduction of OLR is obvious (and interrelated) to the drop of T2M, the reduction in RSR seems to have a more complex mechanism and would merit an additional explanatory sentence*

Reply: To investigate the issue, we conducted additional analyses. The results indicate that the reduction in RSR during the early years of the piControl simulation is primarily driven by ocean adjustment processes and the associated changes in low-level clouds:" While the decrease in OLR is physically consistent with the cooling of T2M, the reduction in RSR is primarily attributed to oceanic adjustment processes. In particular, a cold SST bias (Fig. S3b) induced by the original parameter settings leads to a rapid decline in low-level cloud cover over tropical and subtropical ocean basins—especially in the western Pacific warm pool region and the South Atlantic (Fig. S3c). Most areas of cloud reduction spatially coincide with regions of diminished reflected shortwave radiation (Fig. S3d), a relationship further supported by changes in shortwave cloud forcing (SWCF; Fig. S3e)." (Lines 653-660).

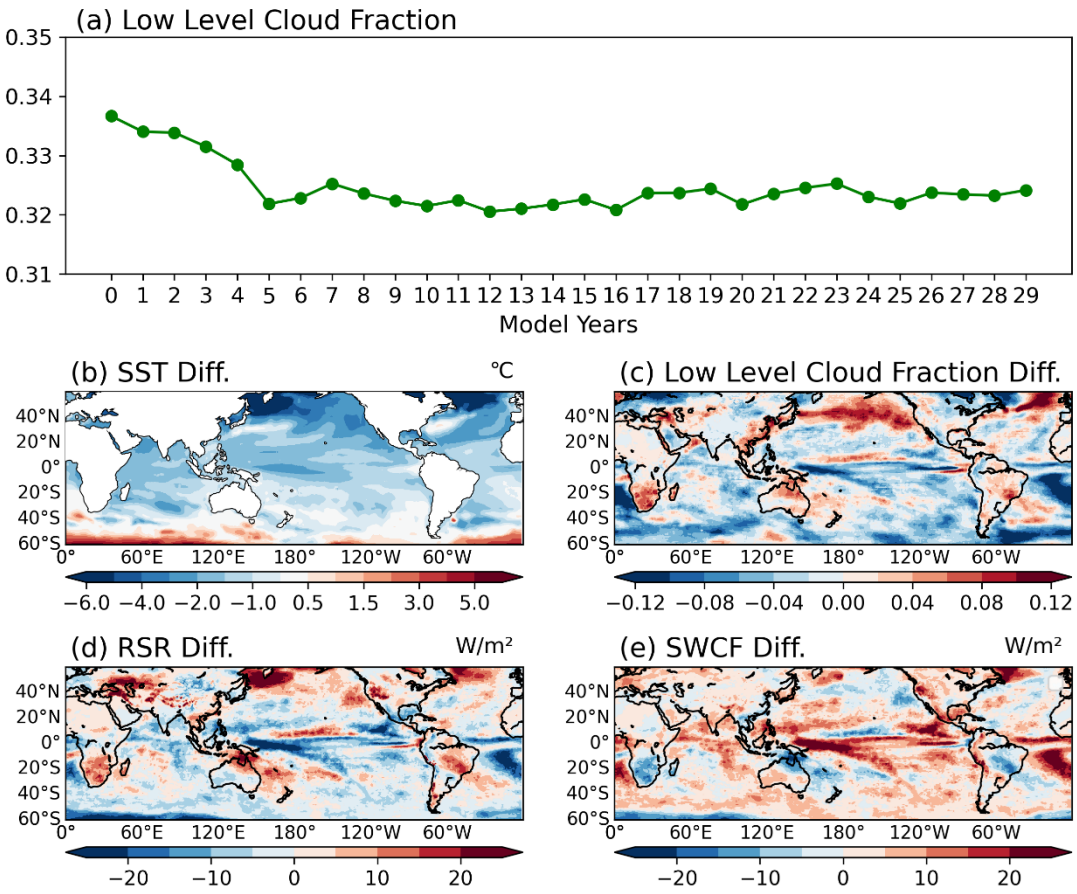

Figure S3. Panel (a) shows the 30-year time series of low-level cloud fraction over the mid- to low-latitude region (60°S–60°N) in the default case. Panels (b)–(e) display the differences between year 6 and year 1 in the piControl run for the default case, including SST (b), low-level cloud fraction (c), RSR (d), and shortwave cloud forcing (SWCF; e).

Comment 14: *L718: a leak of 1.4 W/mw seems quite relevant to me, and , besides being present here, it should have been mentioned earlier in the results when discussing NETFLUX*

Reply: We thank the reviewer for highlighting the issue of energy imbalance. Accordingly, we have included an explicit discussion early in the Results section to explain and discuss the 1.4 W/m² energy leakage:" Further analysis revealed that the relatively large energy imbalance primarily originates from the GAMIL3 atmospheric model, which exhibits a persistent imbalance of approximately 1.4 W/m² in its AMIP configuration—a feature also observed in the piControl runs—due to non-conservation in the dynamical core. This systematic issue is consistent with other atmospheric or coupled models (e.g., up to 1.0 W/m² for CAM6 at 1° resolution (Lauritzen and Williamson, 2019), 1.3 W/m² for FGOALS-g3, and 3.3 W/m² for INM-CM4-8, calculated from Wild, 2020). Notably, this energy leakage

remains stable (±0.1 W/m²) across both default and optimized runs, indicating that the model improvements, such as reduced climate drift, result from genuine parameter tuning rather than compensation for the energy bias. This conclusion is further supported by the coupled model's stabilized energy budget following the spin-up period (Fig. 10)."(Lines 666-676).

Comment 15: *L727: Mention that the primary experiments where 1 -year long AMIP*

Reply: "We have revised the corresponding sentences as follows: "Two primary experiments were conducted using AMIP2011 simulations (2011, with 3-month spin-up): one adjusted 10 parameters and another adjusted 20 parameters. Validation was then performed through extended AMIP2005-2014 and 30-year coupled piControl simulations to assess robustness across timescales." (Lines 883-887).

Comment 16: *L740: the maintained improvement over extended periods is good news given that you tuned on a single year and ignored interannual variability. Could you hypothesise whether (and how much) you would expect a better tuning if you optimize the parameters over several years of AMIP?*

Reply: We appreciate this insightful question regarding the potential benefits of multi-year tuning. A relevant discussion has been added to the manuscript:" while our 1-year optimization produced parameters that remain effective in extended runs (as shown by the AMIP2005–2014 and 30-year piControl validations) and internal variability was explicitly accounted for in the cost function (Eq. 1), including interannual variability—using a longer tuning period like the 5-year approach of Tett et al. (2022)—could further improve results, especially for variables with large interannual variability (e.g., MSLP, Lprecip) and dynamical outputs sensitive to the chosen year. This is supported by Bonnet et al. (2025), who show that short-term tuning works well for physical variables with low interannual variability but multi-year tuning better captures dynamical variability. Based on Bonnet et al. (2025) and our own results—such as the difference observed between 1-year and 10-year simulations for MSLP_TROPICSOCEAN_DGM, which degraded from +20σ to −10σ—we might expect approximately 10–20 % better performance for variables that are particularly sensitive to interannual variability, such as tropical precipitation patterns or extratropical circulation

indices, since a longer tuning period would better sample different climate regimes and reduce sensitivity to single-year anomalies. However, longer tuning greatly increases computational cost—4.2 times higher for 5-year runs. Our current strategy balances efficiency and robustness, but certain metrics like T2M and Lprecip might still benefit from longer tuning. This trade-off warrants further study, particularly where an accurate representation of interannual variability is crucial." (Lines 818-836).

Technical corrections:

Comment 1: *L51 difficult to understand the complete sentence. Perhaps 'carbon cycle or nutrient cycles' would clarify it.*

Reply: Revised to "the coupling of biogeochemical cycles such as the carbon cycle or nutrient cycles with the physical climate system (Erickson et al., 2008)."(Lines 54-56).

Comment 2: *L60: remove 'computational constrains' as it only adds confusion to the sentence.*

Reply: Deleted.

Comment 3: *L239: 'discussed in a later section'. Please state at which specific section.*

Reply: Revised to "will be discussed further in section 2.4."(Line 284).

Comment 4: *L250: listed in the first [instead of last] column of Table 2.*

Comment 5: *L254: listed in the first [instead of last] column of Table 2.*

Reply to the above two: Both revised.

Comment 6: *L273-L277: Break the sentence, it is difficult to follow. L288-L291: assuming there are no typos in the equations, there is inconsistent information in these lines: N is defined twice and differently, and C is defined although missing in the equation.*

Reply: We revised the first sentence to" For the four radiation variables (OLR, OLRC, RSR, and RSRC), uncertainties are based on the estimates from Loeb et al. (2018)" (Lines 325-327).

The formula is correct, and we have revised its explanation as follows "

The cost function is given by:

$$F^2(p) = \frac{1}{N}(S-O)^T C^{-1}(S-O) \quad (2),$$

where *S* is the simulated values; *O* is the target (observed) values; *N* is the number of observations; $(S-O)^T$ is the transpose of the difference between simulated and observed values; C^{-1} is the inverse of the covariance matrix *C* discussed above."(Lines 338-343).

Comment 7: *L357: why not just mention total number of iterations, instead of excluding the first 10?*

Reply: We thank the reviewer for this valuable suggestion. The initial 11 (or 21, depending on the number of tuning parameters) iterations correspond to the mandatory parameter perturbation phase of DFO-LS, during which each parameter is individually perturbed and simulated prior to the optimization process. Since these runs serve as an initialization step rather than part of the iterative optimization, we explicitly distinguished them to avoid overcounting computational costs. For clarity, we have now revised the text in both the Results and Methods sections to report the total number of model evaluations (29 for 10 parameters and 31 for 20 parameters), and we have added a footnote explaining that this count includes the initial perturbation phase. The revision in the Methods section reads as follows:" The optimization process begins with a parameter perturbation phase, in which *K*+1 simulations are conducted: one reference simulation using the initial parameter set, and *K* additional simulations—each perturbing one of the *K* tunable parameters individually—relative to the reference. These initial simulations establish baseline parameter sensitivities and provide finite-difference gradient estimates for the DFO-LS algorithm. The subsequent optimization phase then iteratively modifies parameter values through trust-region managed steps, where each iteration evaluates candidate points, updates local quadratic models of the cost function, and adjusts parameters based on actual versus predicted improvement ratios until convergence criteria are satisfied." (Lines 190-199). The relevant Results section has been revised as follows:" 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. The cost function drops rapidly from about 7.5 to 3.5 during the initial perturbation phase, followed by a slower decline with some fluctuations." (Lines 408-412).

Comment 8: *L403: remove "an.*

Reply: Removed.

Comment 9: *L464: variables.*

Reply: Revised.

Comment 10: *L475: this is less succesfull, in relative terms, than the 10 parameter case.*

Reply: Revised as suggested:" This is less successful, in relative terms, than the 10 parameter case, where 8 variables exhibit reduced or similar bias relative to the default." (Lines 538-540).

Comment 11: *L486: exhibit similar behaviour*

Reply: Revised

Comment 12: *L603: which improvements for which case?*

Reply: Revised to:" with simulated radiation improvements primarily observed in shortwave radiation for the 10-parameter case and in longwave radiation for the 20-parameter case." (Lines 697-699).

Comment 13: *L606: flux of energy towards the ocean, instead of ocean surface flux.*

Reply: Revised.

Comment 14: *L691: a common issue.*

Reply: Revised.

Comment 15: *All figures: larger legends would be good.*

Reply: We have made larger legends for all the figures as suggested.

Comment 16: *Table 2: add units (if they have) to the parameters, as it may help to understand their role.*

Reply: Added.

Comment 17: *Figure 2: the numbers written in the experiment color code are very hard to read. Also, the caption does not explain what they mean, nor the meaning of the vertical dashed lines in b) and c)*

Reply: We have replotted the figure and updated the caption to include further explanation as follows:" Normalized values of tuning parameters for the default and all five optimized cases (a); changes in the cost function values over iterations 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." (Lines 1255-1259). Furthermore, we have clarified in the manuscript that abbreviations such as '10-param.' used in the captions of all relevant figures are explicitly defined in the text, e.g., "This case is denoted as the "10-param." case in the captions of all relevant figures" (Lines 303–304).

Comment 18: *Figure 3: I would rename AMIP@10years by AMIP2005-2014, here and wherever mentioned in the text.*

Reply: Revised as suggested.

Comment 19: *Figure 7: there is a red 'v'.*

Reply: Revised.

Comment 20: *Figure 8: percent instead of precent*

Reply: Revised.

Comment 21: *Figure 12: change colorcode as it uses the same as Figure 7. In Fig 7, however, the numbers in the Table display the actual Jacobians, while here it displays the range between Jacobians. A change of colorcode would help explain that we are not looking at the exact same metric.*

Reply: Changed as suggested.