Response to Reviewer #2

We thank Reviewer #2 for the insightful and constructive evaluation. We have revised the manuscript accordingly and provide point-by-point responses below. Reviewer comments are shown in **bold**, followed by our responses.

Comment 1

This manuscript presents a linear sensitivity–based optimization approach combined with the Nelder–Mead algorithm to improve ENSO simulation in the ICON XPP Earth System Model. The authors conduct a two-stage tuning: first in atmosphere-only mode (21 parameters) and then in fully coupled mode (6 selected parameters), using the CLIVAR ENSO Metrics Package for evaluation. The atmosphere-only optimization achieves roughly 30% reduction in the cost function; the coupled optimization yields improvements in ENSO amplitude, cold tongue bias, phase-locking, and some feedbacks, comparable in magnitude to gains from doubling resolution. The manuscript frankly discusses the unintended global mean warming induced by ENSO-only tuning and its correction via turbulence parameters, and recommends including global constraints (e.g. GMT, AMOC) in future multi-objective optimization.

The methodology is clearly structured, the two-stage design is well justified, and the comparison with CMIP6 and high-resolution ICON XPP is useful. Overall, I recommend MINOR revision: address the points below so that the manuscript is fully consistent and reproducible.

The manuscript contains MANY basic errors in equation/figure cross-references, spelling, grammar, and section numbering that collectively impair readability and raise concerns about the thoroughness of internal review. The authors are strongly urged to conduct a systematic, end-to-end proofread before resubmission. A non-exhaustive list is given below:

- **Incorrect equation references:** The cost function is defined in Section 3.3 as Eq. [6], with Δ_{metric} and Δ_{para} given by Eqs. [7] and [8]. However, in Sections 4.1 and 4.2 the text refers to "Eq. 3" and "Eq. 4".
- **Missing section heading:** Section 5 has subsections 5.1–5.4 but no parent heading (e.g. "5 Results for fully coupled experiments") before 5.1.
- **Incorrect figure reference:** The text "By design, the composite RMSE of the control simulation equals 1.0 (Fig. 2b)" should refer to Fig. 3b.
- **Incorrect map labels:** The "180^{circ}E" label in Figure 7 should be "0^{circ}"; use 0–360^{circ}E for consistency.
- **Spelling errors:** "LIoyd" → "Lloyd" (Line 67 and 95), "fileds" → "fields" (Line 131), "re-girded" → "regridded" (Line 134), and "metrices" → "metrics" (Lines 189, 190, and 192).
- **Inconsistent parameter names:** unified the parameter name to "tune_entrorg" (including the Section 3.3 text) for consistency with Table 1.

Response: Thank you very much for this thorough and constructive review, and especially for highlighting the consistency and reproducibility issues. We agree with your assessment and have carried out a systematic end-to-end proofread of the full manuscript.

We revised all points listed in this comment as follows:

- **Equation cross-references:** corrected incorrect references in Sections 4.1 and 4.2 so the cost-function discussion consistently cites Eqs. [6], [7], and [8] from Section 3.3.
- **Section hierarchy:** added a parent heading for Section 5 before subsection 5.1 ("Results for fully coupled experiments"), so subsection numbering and structure are now consistent.
- **Figure cross-reference:** corrected "Fig. 2b" to "Fig. 3b" in the sentence about the composite RMSE of the control simulation.
- **Figure 7 longitude labels:** corrected the map labeling issue (including replacing the incorrect "180°E" label with "0° where appropriate) and harmonized longitude labeling to the 0–360°E convention for consistency with the rest of the manuscript.
- **Spelling and proofreading corrections:** corrected "LIoyd" to "Lloyd", "fileds" to "fields", "regirded" to "regridded", and "metrices" to "metrics", and performed an additional full-manuscript language and formatting cleanup to reduce similar issues.
- **Parameter-name consistency:** standardized parameter naming throughout, including Section 3.3, to match Table 1.

These revisions substantially improve internal consistency, readability, and reproducibility of the manuscript. We appreciate this important quality-control feedback.

In addition to the originally listed examples, our full recheck identified and corrected several further basic issues:

- In Section 3.1, we corrected additional residual grammar beyond the reviewer list, for example revising the teleconnection-metric sentence from "which will therefore not used" to "which will therefore not be used".
- In Section 3.4, we corrected further wording and notation inconsistencies, including changing "as shown as" to "as shown by" and aligning the normalized parameter notation in Eq. [9] from Δnp_i to Δnp_k.
- In Section 5.1, we removed the duplicated phrase "fully coupled fully coupled experiments", which had remained in the original draft.
- In the fully coupled sensitivity discussion, we corrected additional grammar issues such as the subject–verb agreement in "The results suggest ...", which was inconsistent in the original manuscript.
- We also cross-checked captions, section text, and parameter notation to ensure that figure names, symbols, and parameter labels are used consistently after revision.

Comment 2

For data grid resolution, Section 2.1 states that the model data and observational data are regridded to the same $1^\circ \times 1^\circ$ global grid, while Section 2.2 specifies that the model runs at 160 km atmosphere ($\sim 1.4^\circ$) and 40 km ocean ($\sim 0.36^\circ$). This creates confusion. (1) What are the native grid resolutions of each observational dataset and of the ICON model output, respectively? (2) What regridding/interpolation

method was used to bring all data onto the 1°×1° grid? (3) Could interpolating the 160 km atmosphere (∼1.4°) to a finer 1° grid introduce artificial smoothness in local metrics such as the meridional precipitation structure? (4) Could coarsening the 40 km ocean output to 1° obscure sub-degree small-scale ocean features (e.g., tropical instability waves, sharp SST fronts) that may be relevant to some ENSO feedbacks?

Response: Thank you for this careful and important comment. We agree that the original manuscript did not explain grid resolutions and regridding procedures with sufficient precision. We have now revised the manuscript to explicitly address all four points.

- **(1) Native resolutions:** We now explicitly state native resolutions: GPCPv2.3 precipitation (2.5°×2.5°), TropFlux variables (SST, wind stress, heat flux; 1°×1°), and GODAS SSH (1°×1/3°). ICON XPP output uses native unstructured grids: atmosphere at R2B4 (∼160 km, about 1.4°) and ocean at R2B6 (∼40 km, about 0.36°).
- **(2) Regridding method:** All model and observational fields are remapped to a common 1°×1° latitude–longitude grid using first-order conservative remapping. We also now state explicitly that this is consistent with the ENSO metrics overview implementation, where data are interpolated onto a generic 1° latitude × 1° longitude grid (Planton et al., 2021).
- **(3) Atmosphere ∼1.4° to 1°:** We agree this interpolation does not increase effective resolution and can introduce some smoothing. We now clarify that the ENSO diagnostics used here focus primarily on large-scale structures (e.g., zonal means, equatorial averages, seasonal cycles, basin-scale feedbacks), for which this remapping choice is appropriate (Planton et al., 2021).
- **(4) Ocean 0.36° to 1°:** We agree that coarsening may filter sub-degree features (e.g., tropical instability waves, sharp fronts). We now explicitly acknowledge this limitation and clarify that the present evaluation targets basin-scale ENSO metrics rather than small-scale ocean processes.

In summary, we clarified native dataset/model resolutions, explicitly documented the remapping method (first-order conservative remapping), and added a concise discussion of potential smoothing/scale-filtering effects and their relevance to the large-scale ENSO metrics analyzed in this study.

Comment 3

For experiment period range, Section 2.1 states that the period for all observational reference data is 1980–2018, while Section 2.3 indicates that the AO experiments cover 1979–1997. The first year of the AO experiments (1979) has no corresponding observational reference. (1) When calculating ENSO metrics, is the AO model output compared with observations using only 1980–1997, or the full 1979–1997? If the latter, what is the source of observations for 1979? (2) The AO experiments do not cover the 1998–2018 period, which includes several strong ENSO events. Does this affect the representativeness of the sensitivity estimates?

Response: Thank you for this important comment. We agree that the temporal setup should be stated more clearly.

(1) In the atmosphere-only (AO) configuration, ENSO metrics in this study are computed from the AO simulation period 1979–1997. To test whether inclusion of 1979 affects the diagnostics,

we compared ENSO metrics from 1979–1997 and 1980–1997 for the AO model reference configuration (Figure R2-1). The two metric sets are nearly indistinguishable, indicating that including 1979 has negligible impact on our conclusions.

(2) We agree that 1979–1997 does not include later strong ENSO events. The original AO ensemble length was limited by a restart issue in the AMIP workflow, so we used the longest continuous period available at that time. This restart issue has now been fixed, and we extended the AO reference run to 2014 (the maximum end year under our CMIP6 historical forcing setup). A sensitivity comparison among 1979–1997, 1980–1997, and 1979–2014 again shows only very small differences in ENSO metrics (Figure R2-1).

These tests demonstrate that, for the atmosphere-only reference configuration, the ENSO metrics used here are dominated by climatological and large-scale mean-state characteristics; therefore, both inclusion of 1979 and the shorter AO period have negligible influence on the inferred sensitivity patterns. We also added these discussions into the main text.

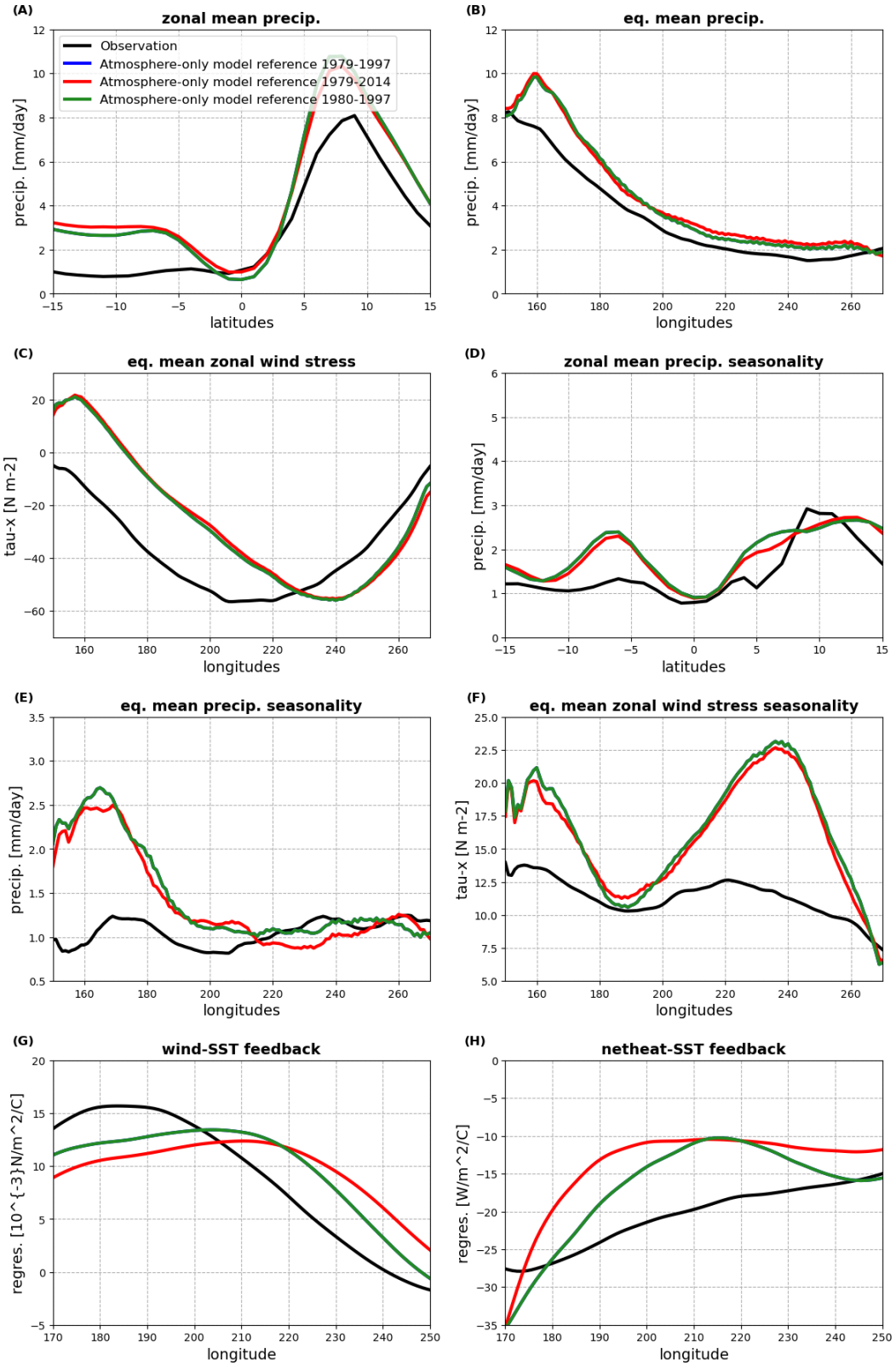

Figure R2-1. Sensitivity check of ENSO metrics for the atmosphere-only model reference configuration using different analysis periods (1979–1997, 1980–1997, and 1979–2014). The metric curves are nearly overlapping, indicating weak sensitivity to period choice.

Comment 4

In Table 1, multiple optimized values fall outside the stated ranges. Eq. [10] and the accompanying text imply that parameter combinations violating physical bounds are penalized and discarded, yet many parameters exceed the stated ranges.

Response: Thank you for this careful observation. We agree that the distinction between parameter ranges and parameter constraints should be made more explicit.

The key point is that two different constraint concepts are used in this study:

(1) **Physical bounds (hard constraints):** Eq. [10] is applied to fundamental physical limits (for example, positivity for selected parameters). Candidate parameter combinations that violate these physical bounds are excluded from the optimization search.

(2) **Perturbation ranges (not hard constraints):** The ranges listed in Table 1 are the perturbation intervals used in the initial sensitivity experiments. They are not enforced as strict bounds during the subsequent optimization step.

Therefore, optimized values outside the listed Table 1 ranges mainly occur when the optimum lies near or slightly beyond the edge of the initial sampling space. In practice, these deviations are moderate and do not lead to physically unrealistic model behavior. We have revised the manuscript to clarify that Table 1 ranges are perturbation intervals used for sensitivity estimation, and that optimization applies a soft regularization constraint for parameter deviation while Eq. [10] still enforces hard physical bounds.

Comment 5

Eq. [11] approximates the metric bias δ_m by a linear combination of parameter sensitivities and is central to the computational efficiency of the scheme. However, ENSO is a complex air-sea coupled system with highly nonlinear characteristics. (1) Why is it reasonable to use a linear superposition approximation in such a complex nonlinear system? It seems to lack a physical or dynamical explanation for this basic assumption. (2) Over what range of parameter changes is this approximation expected to hold? (3) What is the impact of neglecting nonlinear and cross-parameter terms (interaction terms) on the optimization results?

Response: Thank you for this important and insightful comment. We agree that the use of a linear superposition approximation in a highly nonlinear system such as ENSO requires careful justification. We have revised the manuscript to clarify the underlying assumptions, validity range, and limitations of this approach.

(1) **Justification of the linear approximation.** The linear superposition used in Eq. (11) can be interpreted as a first-order Taylor expansion of the model response around a reference (control) state. Specifically, the sensitivity estimates approximate how individual ENSO-related variables respond locally to parameter perturbations. While ENSO dynamics are inherently nonlinear, the approximation is applied to small perturbations in model parameters, rather than to the full system evolution. As such, the method captures the leading-order response of the model to parameter changes, which is a standard approach in sensitivity analysis.

Importantly, the approximation is not intended to represent the full nonlinear dynamics of ENSO, but rather to provide a local, first-order estimate of how parameter perturbations affect ENSO-related metrics.

(2) **Valid range of the approximation.** The validity of the linear approximation is expected to hold within a local neighborhood of the reference parameter set, where parameter perturbations remain moderate. In this study, the perturbation ranges used to estimate sensitivities are relatively small and guided by physically plausible parameter intervals. This ensures that the linear approximation remains applicable and avoids strongly nonlinear regimes.

(3) **Impact of neglecting nonlinear and interaction terms.** We acknowledge that the linear superposition neglects higher-order nonlinear effects and cross-parameter interactions. As a result, the method may not fully capture complex parameter dependencies or synergistic effects between parameters. This can introduce approximation errors in the estimated metric response.

However, in practice, the optimization results show consistent improvements in ENSO metrics, indicating that the first-order approximation captures the dominant sensitivities relevant for parameter tuning. Moreover, the use of ensemble-based sensitivity estimates helps to partially account for variability and reduces the impact of noise.

We have added discussion in the revised manuscript to explicitly acknowledge these limitations and to clarify that the method should be interpreted as a computationally efficient first-order approximation, rather than a fully nonlinear optimization framework. Extending the approach to include nonlinear interactions and nonlinear optimization is an important direction for future work.

Comment 6

The "RMS threshold of 0.2" is used to select parameters with "significant" impact on ENSO metrics. The text states it corresponds to roughly 20% of the control-run bias amplitude. (1) Was this value chosen from a break in the sensitivity distribution or from a signal-to-noise criterion? (2) Has this empirical threshold been used in previous studies?

Response: Thank you for this thoughtful comment. We agree that the choice of the RMS threshold requires clarification.

(1) **Basis for the threshold selection.** The RMS threshold of 0.2 was chosen as a pragmatic criterion to identify parameters with non-negligible influence on ENSO metrics, rather than from a strict breakpoint analysis or a formal signal-to-noise cutoff. The value corresponds to approximately 20% of the control-run bias amplitude, which provides a physically interpretable reference scale for screening influential parameters.

(2) **Relation to previous studies.** To our knowledge, this exact numerical value (0.2) is not a standard universal threshold in previous studies. However, using relative or normalized criteria to screen influential parameters is common in climate-model tuning and sensitivity analysis (e.g., Hourdin et al., 2017; Williamson et al., 2013; Saltelli et al., 2008).

Therefore, our threshold should be interpreted as a transparent and practical screening criterion rather than a theoretically optimal universal cutoff. We have revised the manuscript to make this point explicit.

References used to support this clarification:

- Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., Rauser, F., Roehrig, R., Svensson, G., Watanabe, M., and Williamson, D.: *The art and science of climate model tuning.* Bulletin of the American Meteorological Society, 98(3), 589–602, 2017. https://doi.org/10.1175/BAMS-D-15-00135.1

- Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., and Yamazaki, K.: *History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble.* Climate Dynamics, 41, 1703–1729, 2013. https://doi.org/10.1007/s00382-013-1896-4
- Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S.: *Global Sensitivity Analysis: The Primer.* John Wiley & Sons, Chichester, UK, 2008. https://doi.org/10.1002/9780470725184

Comment 7

The six parameters carried into the FC optimization were selected based on their AO sensitivity ranking. Given that AO and FC sensitivities differ substantially in both magnitude and sign, the AO ranking may not reliably reflect parameter importance in the coupled system. (1) Could parameters with weak AO sensitivities nonetheless have significant impacts in the coupled configuration? (2) What are the potential consequences of this selection strategy for the final coupled optimization?

Response: Thank you for this insightful comment. We agree that parameter sensitivities can differ between atmosphere-only (AO) and fully coupled (FC) configurations, including differences in both magnitude and sign due to ocean–atmosphere feedbacks.

(1) **Could weak AO parameters still matter in FC?** Yes, this is possible. Parameters with weak AO sensitivity can become more influential in FC through coupled feedback pathways that are absent in AO experiments.

(2) **Potential consequences and rationale.** The AO-based ranking is used as a first-order screening strategy rather than a definitive ranking of FC importance. This choice is motivated by computational feasibility: exhaustive FC sensitivity exploration over the full parameter space is currently too expensive. The main consequence is that some parameters with weak AO signal but stronger FC influence could be missed in the reduced FC search space. At the same time, AO screening efficiently identifies parameters with strong direct atmospheric influence on ENSO-relevant fields, and the subsequent FC optimization still yields clear ENSO improvements.

We have revised the manuscript to explicitly state this limitation and to clarify that AO-based pre-selection is a practical dimensionality-reduction step. More comprehensive FC-side sensitivity exploration, including potentially FC-specific parameters not highlighted by AO ranking, is an important direction for future work.

Comment 8

The Nelder–Mead method can converge to local minima. Was the optimization run from a single initial point or from multiple starting points?

Response: Thank you for this important comment. We agree that Nelder–Mead is a local optimization method and can converge to local minima depending on initialization.

In this study, the optimization was initialized from a single starting point: the control parameter set. We chose this baseline because it is physically calibrated and numerically stable. Our objective here is to obtain a robust local improvement of ENSO-related metrics relative to this reference state, rather than to identify a global optimum over the full high-dimensional parameter space.

We acknowledge that different initial points could, in principle, lead to different local optima. Exploring multiple initializations and/or global optimization strategies is an important direction for future work, and we now state this explicitly in the manuscript.