

Response to the editor review:

We thank the Editor for the careful assessment of our revised manuscript and for the positive evaluation of the revisions. We have addressed all of the remaining technical concerns raised in the report. In particular, we have (i) added a brief discussion of the contrasting performance of the iLOVECLIM simulations between the global and Indian Ocean evaluations, and (ii) revised 5 figures to ensure that the observational dataset used for comparison is strictly identical across all panels and to clarify the treatment of simulation outliers. We have also made additional minor changes to improve the consistency and clarity of the figures and text. Our detailed responses to each comment are provided below, in purple.

It is remarkable that the iLOVECLIM simulations show the most skill in the evaluation of the global LGM and PI fields (section 3), but the lowest skill in the Indian Ocean (section 4). Do you have an explanation? At the very least this should be briefly mentioned and discussed.

We thank the Editor for pointing this out. We agree that this apparent contrast in iLOVECLIM performance between the global evaluation and the Indian Ocean deserves clarification. We have therefore added a brief discussion in Section 4 to highlight this difference and discuss possible reasons for the lower skill of the iLOVECLIM simulations in the Indian Ocean compared with their performance at the global scale.

Was added L.661:

“The particularly poor performance of iLOVECLIM in the North Indian Ocean likely contributes to the contrast between its global and regional skill. Because the global evaluation combines information across the entire ocean domain, strong performance in other regions may compensate for poorer agreement in the Indian Ocean.”

You have added several figures with the histograms (for example Figure 6). I am confused that the data IQR is different in each of the panels. Shouldn't the data histogram (grey) be identical in all panels given that model performance is evaluated against the same data set? Or are they different because of the Monte-Carlo procedure? Please either explain why they are different, or fix the figures. It appears this is true for all the histogram figures. Apologies if there is an explanation provided for this that I missed.

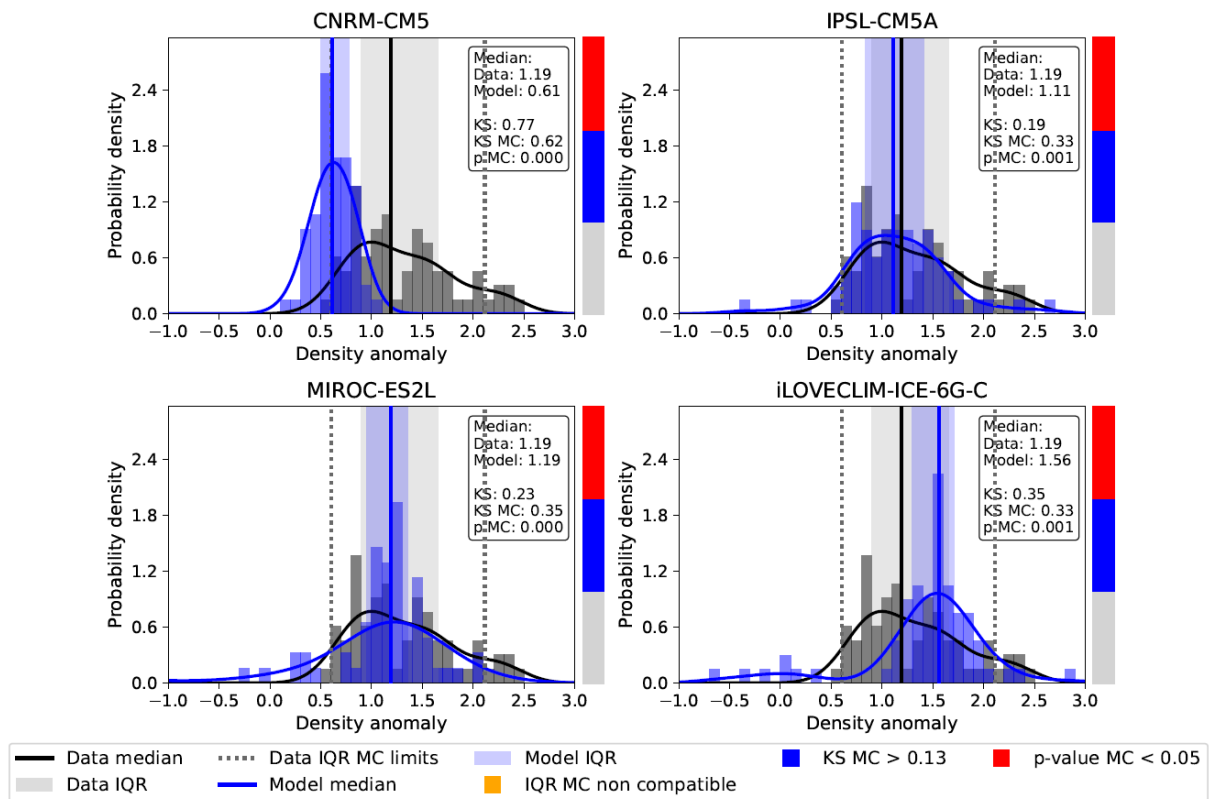
We thank the Editor for identifying this issue. We agree that the observational distribution, and therefore its IQR, should be identical across all panels of a given histogram figure. The differences in the observational IQR resulted from the fact that the comparisons did not always use exactly the same observational points: some observational locations did not correspond to valid grid points in all model simulations, resulting in slightly different observational subsets being used for different simulations.

We have corrected this by retaining only the observational points for which a valid value is available in all model simulations included in the comparison. The same observational dataset is therefore now used for every panel, ensuring that the observational histogram and IQR are identical across panels. This correction has been applied consistently to all histogram figures.

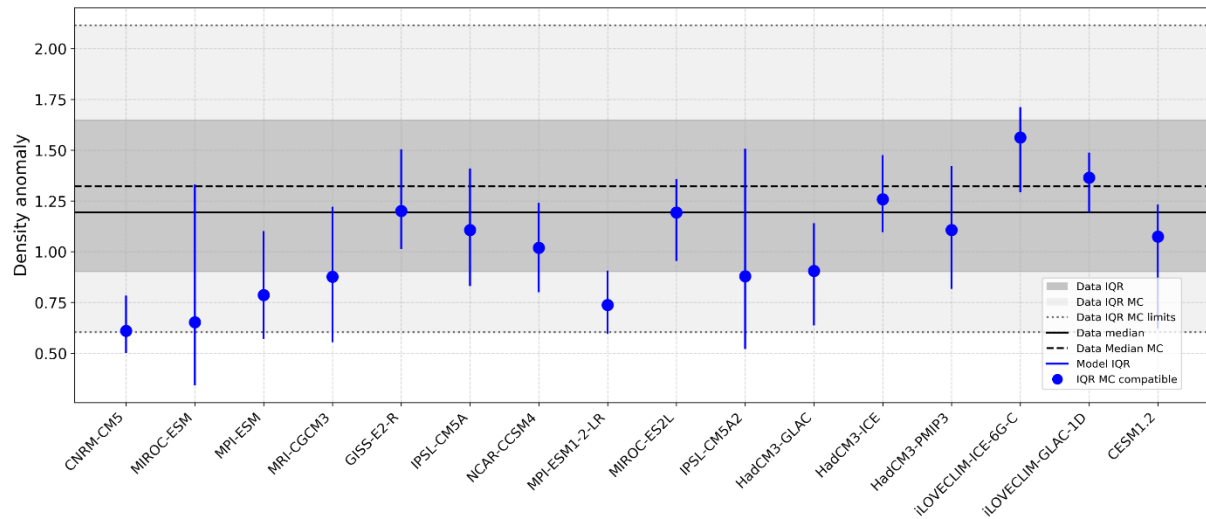
In addition, we have revised the treatment of the model distributions to retain all simulation values, including outliers. No observational outliers were identified and removed. For visual clarity, the histogram x-axes are truncated to a common range, but all simulation values, including those falling outside the displayed range, are retained in the statistical calculations. Consequently, in figures 7, S3 b) and S4 b), the model median may not necessarily lie at the centre of the displayed IQR when outlying simulation values are included.

We thank the Editor for bringing this inconsistency to our attention, which allowed us to improve the consistency of the comparison and presentation of the results.

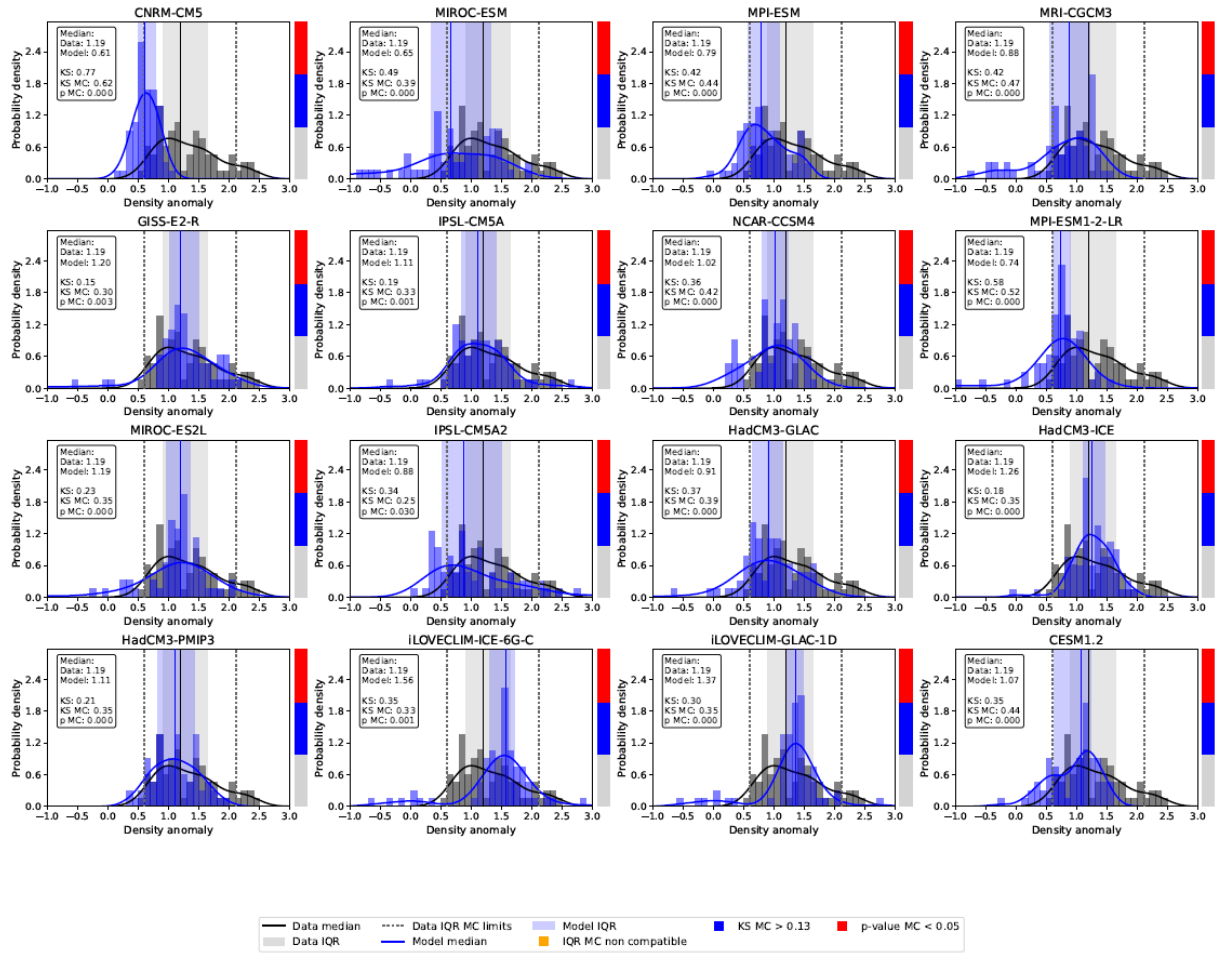
Here are the revised figures:



Revised Figure 6: Distribution histograms of surface density anomalies (LGM-PI, kg/m³) for an example of four model simulations. Density reconstructions are shown in black, model simulations in blue. Kernel Density Estimates (KDEs) illustrate the central tendency and overall shape of the distributions. Vertical lines indicate the median of each distribution, and shaded envelopes represent the interquartile ranges (IQRs), providing a measure of data spread that is independent of extreme values. Grey dotted lines indicate the limits of the Monte Carlo propagated reconstruction IQR obtained after accounting for the 95% confidence intervals associated with the density reconstructions. The histograms display the frequency of values, complementing the KDEs and IQRs to give an integrated view of distribution characteristics. Reconstructions uncertainties are explicitly propagated through the Monte Carlo procedure. Colored indicators on the right side of each panel summarize the outcomes of the three comparison metrics. A red flag indicates a Monte Carlo-derived two-sample KS test with $p < 0.05$, a blue flag indicates a Monte Carlo-derived KS statistic exceeding the global critical threshold ($KS > 0.13$), and an orange flag indicates non-overlapping Monte Carlo propagated IQRs between reconstructed and simulated distributions.

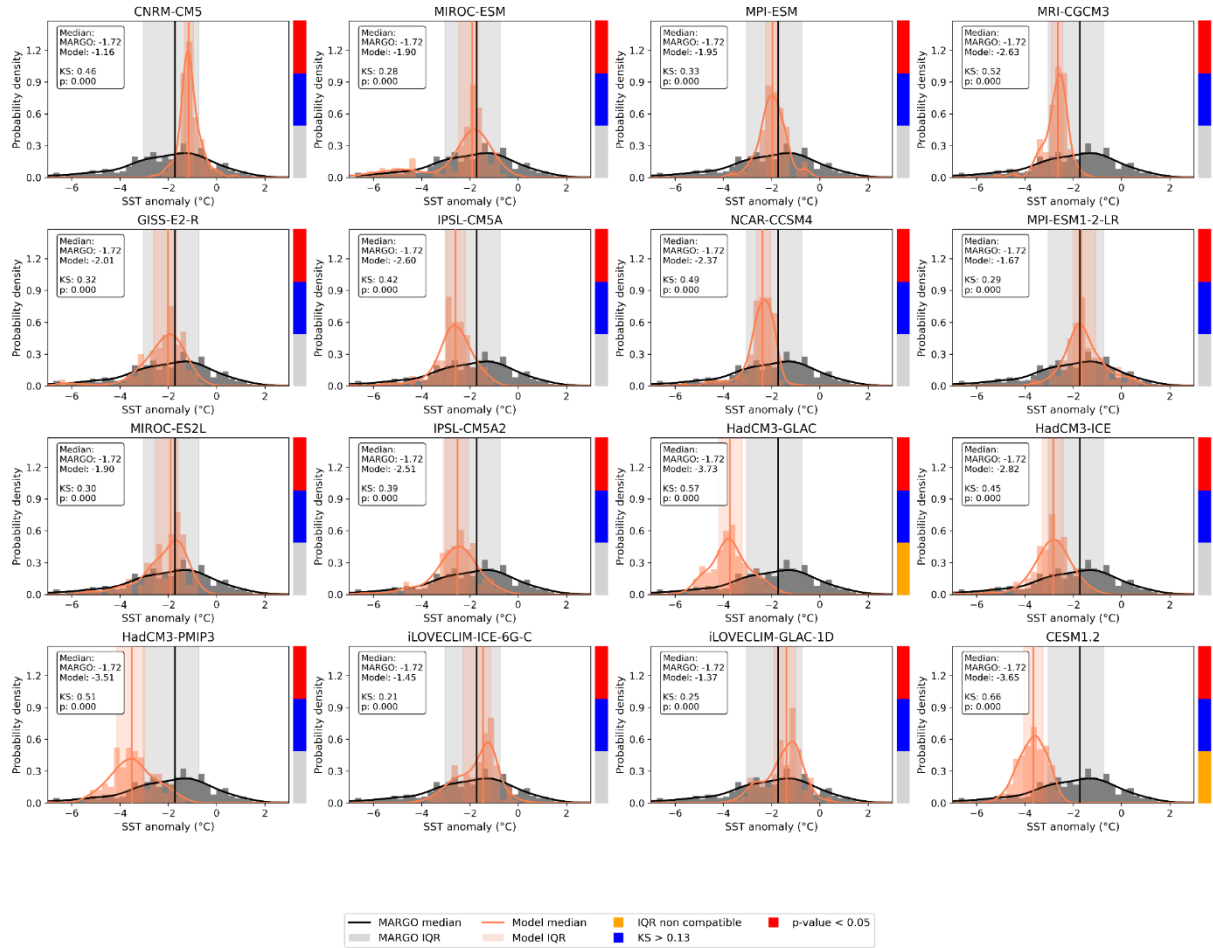


Revised Figure 7: The dark grey band indicates the reconstructed IQR without considering uncertainties, while the light grey band and grey dotted lines indicate the Monte Carlo propagated reconstruction IQR after accounting for reconstruction uncertainty. The solid black line marks the reconstructed median and the dashed black line the Monte Carlo median. For each model simulation, the blue vertical bars represent the modelled IQR, while the circular markers represent the model median. Blue markers indicate model simulations with IQR values consistent with the Monte Carlo propagated reconstruction IQR.

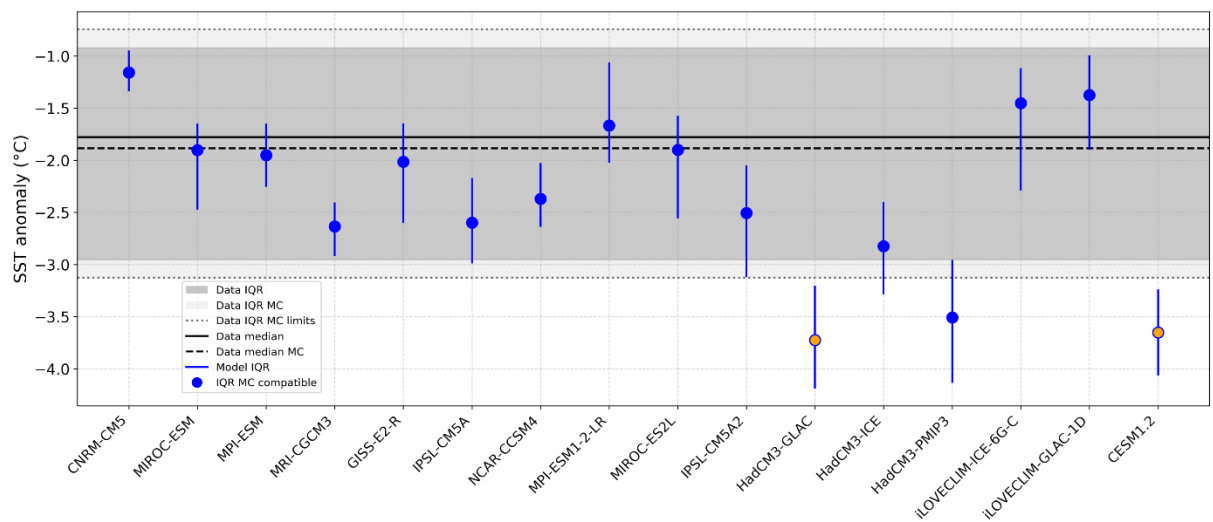


Revised Figure C1: Distribution histograms of surface density anomalies (LGM-PI, kg/m³) at the global scale. Density reconstructions are shown in black, model simulations in blue. Kernel Density Estimates (KDEs) illustrate the central tendency and overall shape of the distributions. Vertical lines indicate the median of each distribution, and shaded envelopes represent the interquartile ranges (IQRs), providing a measure of data spread that is independent of extreme values. Grey dotted lines indicate the limits of the Monte Carlo propagated reconstruction IQR obtained after accounting for the 95% confidence intervals associated with the density reconstructions. The histograms display the frequency of values, complementing the KDEs and IQRs to give an integrated view of distribution characteristics. Reconstructions uncertainties are explicitly propagated through the Monte Carlo procedure. Colored indicators on the right side of each panel summarize the outcomes of the three comparison metrics. A red flag indicates a Monte Carlo-derived two-sample KS test with $p < 0.05$, a blue flag indicates a Monte Carlo-derived KS statistic exceeding the global critical threshold ($KS > 0.13$), and an orange flag indicates non-overlapping Monte Carlo propagated IQRs between reconstructed and simulated distributions.

a)



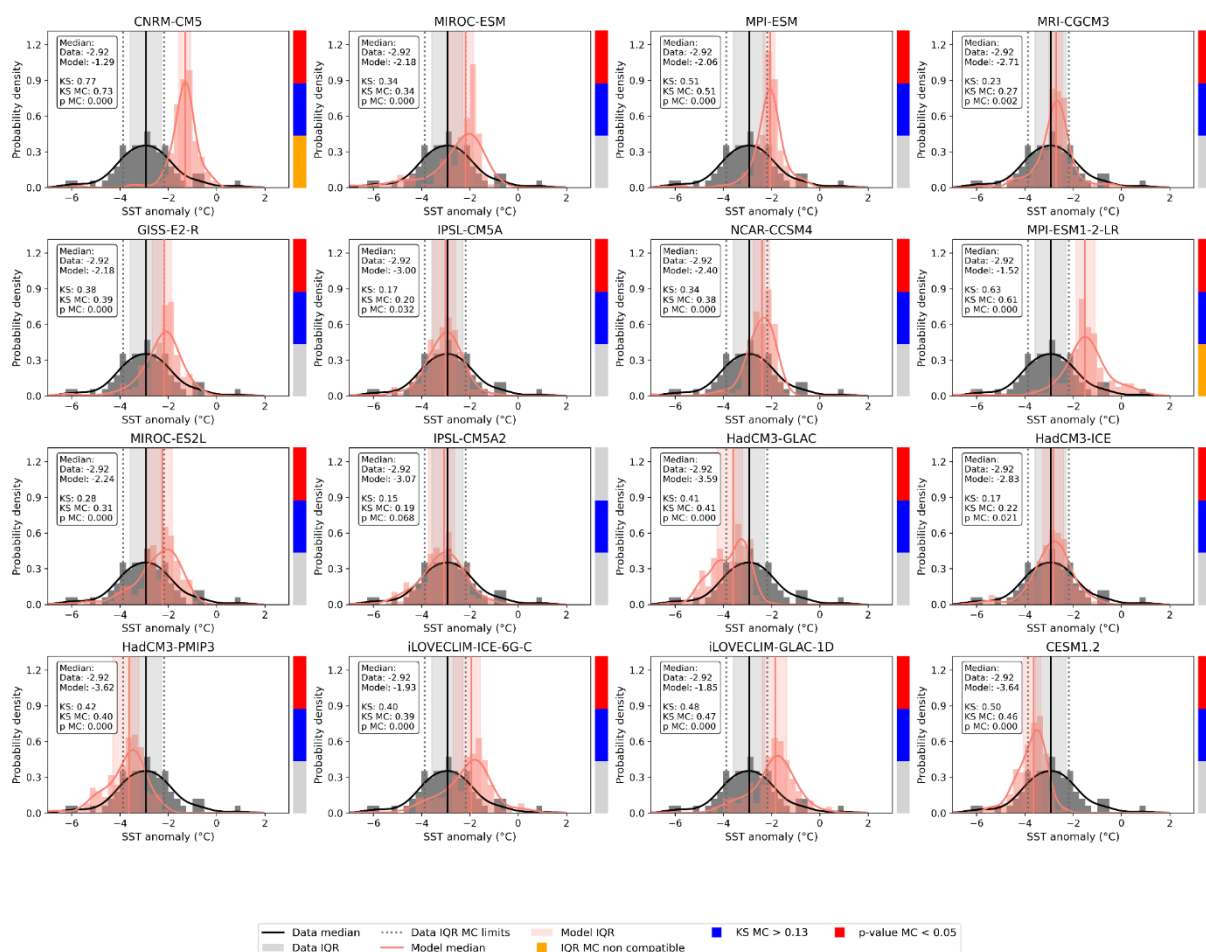
b)



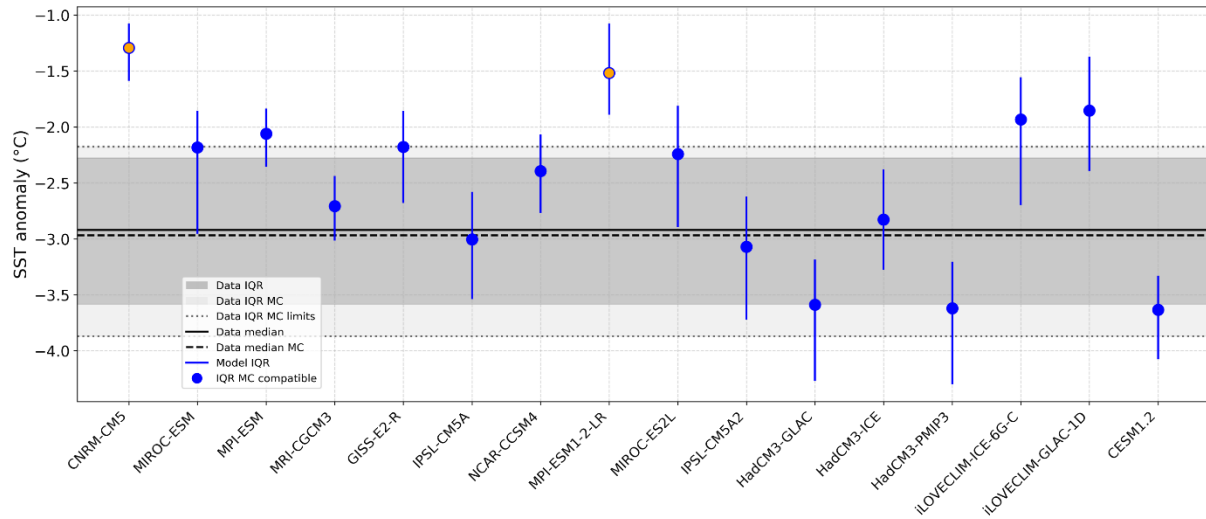
Revised Figure S3: (a) Distribution histograms of SST anomalies (LGM-PI, °C) for each ocean basin. SST reconstructions from the MARGO database (MARGO project, 2009) are shown in black, and model simulations in orange. Kernel Density Estimates (KDEs) illustrate the central tendency and shape of the distributions. Vertical lines indicate medians, and shaded envelopes represent interquartile ranges (IQRs), capturing spread independent of outliers. Histograms show frequency and complement the KDEs. Reconstruction uncertainties are explicitly propagated through the Monte Carlo procedure. Grey dotted lines indicate the limits of the Monte Carlo propagated reconstruction IQR obtained after accounting for the 95 % confidence intervals associated with the SST reconstructions. Colored indicators at the right of each panel show whether any of the following criteria are not met: Monte Carlo-derived p-value < 0.05, Monte Carlo-derived KS statistic ≥ 0.13 , and non-

overlapping Monte Carlo propagated IQRs between reconstructed and simulated distributions. A colored flag indicates a failed criterion. **(b)** Comparison of observed and simulated SST anomaly statistics at the global scale. The dark grey band indicates the reconstructed IQR, while the light grey band and grey dotted lines indicate the Monte Carlo propagated reconstruction IQR after accounting for reconstruction uncertainty. The solid black line marks the reconstructed median and the dashed black line the Monte Carlo median. For each model simulation, the blue vertical bars represent the modelled IQR, while the circular markers represent the model median. Blue markers indicate model simulations with IQR values consistent with the Monte Carlo propagated reconstruction IQR, whereas orange markers highlight models with non-overlapping Monte Carlo propagated reconstruction IQRs.

a)



b)



Revised Figure S4: (a) Distribution histograms of SST anomalies (LGM–PI, °C) for each ocean basin. SST reconstructions from the Tierney database (Tierney et al., 2020) are shown in black, and model simulations in orange. Kernel Density Estimates (KDEs) illustrate the central tendency and shape of the distributions. Vertical lines indicate medians, and shaded envelopes represent interquartile ranges (IQRs), capturing spread independent of outliers. Histograms show frequency and complement the KDEs. Reconstruction uncertainties are explicitly propagated through the Monte Carlo procedure. Grey dotted lines indicate the limits of the Monte Carlo propagated reconstruction IQR obtained after accounting for the 95 % confidence intervals associated with the SST reconstructions. Colored indicators at the right of each panel show whether any of the following criteria are not met: Monte Carlo-derived p-value < 0.05, Monte Carlo-derived KS statistic ≥ 0.13 , and non-overlapping Monte Carlo propagated IQRs between reconstructed and simulated distributions. A colored flag indicates a failed criterion. **(b)** Comparison of observed and simulated SST anomaly statistics at the global scale. The dark grey band indicates the reconstructed IQR, while the light grey band and grey dotted lines indicate the Monte Carlo propagated reconstruction IQR after accounting for reconstruction uncertainty. The solid black line marks the reconstructed median and the dashed black line the Monte Carlo median. For each model simulation, the blue vertical bars represent the modelled IQR, while the circular markers represent the model median. Blue markers indicate model simulations with IQR values consistent with the Monte Carlo propagated reconstruction IQR, whereas orange markers highlight models with non-overlapping Monte Carlo propagated reconstruction IQRs.

Line-by-line comments:

Line 57: Remove “and the two iLOVECLIM simulations consistently perform more poorly than the rest of the ensemble”: this seems too detailed for the abstract, readers have not been introduced to the models yet. Also, they seem to agree well in other respects (see for example line 555).

Thank you for this comment. We have removed this part of the sentence from the abstract.

Line 267: incorrect grammar. I suggest: To further compare data and model simulations we introduce other metrics to reduce overreliance on regression and correlation.

Thank you for the suggestion. We have replaced the original sentence with the proposed one.

Line 547, 642, 677, figure captions: For the reader less familiar with the various statistical methods, please repeat what values constitute a good fit for each of these criteria (“Values close to 1 indicate high skill” etc.).

Thank you for this suggestion. We have added explanations of the optimal values and interpretation of each metric to the relevant figure captions (Figs. 9–11) :

Figure 9: Global model simulations performance for sea surface density during the LGM and PI periods. Heatmaps show four complementary metrics calculated from model–data comparisons at reconstruction grid cells common to all simulations: bias, root mean square error (RMSE), Nash–Sutcliffe efficiency (NSE), and Kling–Gupta efficiency (KGE). **For bias and RMSE, values close to 0 indicate better agreement, whereas for NSE and KGE, values close to 1 indicate higher skill. Negative NSE values indicate poor model performance.** Models are ranked according to their LGM KGE values. For each metric, LGM and PI are shown side by side using the same colour scale, allowing direct comparison of model performance between climatic periods.

Figure 10: Simulations performance for seawater density in the North Indian and South Indian oceans during the Last Glacial Maximum (LGM). Heatmaps show four complementary performance metrics calculated from model–data comparisons at proxy reconstruction sites. Models are ranked according to their weighted mean Kling–Gupta Efficiency (KGE), averaged across both basins. The metrics shown are: mean absolute bias ($|\text{Bias}|$), root mean square error (RMSE), Nash–Sutcliffe efficiency (NSE), and Kling–Gupta efficiency (KGE). For each simulation, the weighted mean values were computed using the number of valid grid points available within each basin. **Lower values of $|\text{Bias}|$ and RMSE indicate better agreement with the reconstructions, with values close to 0 being optimal, whereas higher values of NSE and KGE indicate better model performance, with an optimum value of 1. Negative NSE values indicate poor model performance.**

Figure 11: Simulations performance for seawater density in the North Indian and South Indian oceans during the pre-industrial period (PI). Heatmaps show four complementary performance metrics calculated from model–data comparisons at proxy reconstruction sites. Models are ranked according to their weighted mean Kling–Gupta Efficiency (KGE), averaged across both basins. The metrics shown are: mean absolute bias ($|\text{Bias}|$), root mean square error (RMSE), Nash–Sutcliffe efficiency (NSE), and Kling–Gupta efficiency (KGE). For each simulation, the weighted mean values were computed using the number of valid grid points available within each basin. **Lower values of $|\text{Bias}|$ and RMSE indicate better agreement with the reconstructions, with values close to 0 being optimal, whereas higher values of NSE and KGE indicate better model performance, with an optimum value of 1. Negative NSE values indicate poor model performance.**

Line 565, 581: change “least successful” to “least skillful” or “poorer performing”

Done. We have replaced these expressions with “least skillful” and “poorer performing”, respectively.

L.565:

In contrast, HadCM3-GLAC, HadCM3-ICE, HadCM3-PMIP3 and MIROC-ESM consistently rank among the **least skillful** simulations.

L.582:

The HadCM3 simulations and MIROC-ESM remain among the **poorer performing** during the PI, exhibiting the largest RMSE values together with comparatively low NSE and KGE values, although their performance is generally improved relative to the LGM.

Line 577: change “performances” to “performance”

Done. “Performances” has been changed to “performance”.

Line 612: place period after “dynamics” to end the sentence. Start next sentence with “Previous studies...” and Line 613: change “enlightening” to “revealing”

Thank you for the suggestions. We have revised the sentence accordingly:

“This choice is motivated by the central role of the Indian Ocean in tropical–extratropical interactions and its importance for major hydroclimate processes such as the monsoon and Walker circulation dynamics. Previous studies have already conducted data–model comparisons in this region, revealing its complexity.”

Line 631: “model simulation skill”

Done. We have changed “model simulations skill” to “model simulation skill”.

Line 640: Maybe point out that this is unexpected given the greater skill of the iLOVECLIM model in the PI and LGM evaluations presented in the previous section.

Thank you for pointing this out. We have added a sentence highlighting the contrast between the global and regional evaluations:

“In contrast, the two iLOVECLIM simulations clearly stand apart, exhibiting markedly lower KGE values and substantially poorer agreement with the reconstructions. **This contrasts with their strong performance in the global LGM and PI evaluations (Fig. 9), where the two iLOVECLIM simulations ranked among the best-performing models. This discrepancy highlights the regional dependence of model–data agreement and suggests that good global performance can mask substantial regional differences.**”

Line 671 (During the PI (Fig. 11), it remains very similar compare to the LGM). Please rewrite. Incorrect grammar, and unclear what is meant by “it”.

Thank you for the comment. We have revised the sentence to:

“During the PI (Fig. 11), **the results remain very similar to those obtained for the LGM.**”

Line 859: what do you mean by “to resume”? Do you mean “to summarize”?

Yes, we meant “to summarize”. We have corrected the sentence accordingly:

“**To summarize**, our results confirm a strong cooling in the eastern Indian Ocean contrasted with milder cooling in the western basin, leading to a reduction of the zonal SST gradient compared to preindustrial conditions (DiNezio et al., 2018).”