

Response to reviewers

We thank the Editor and both Reviewers for their thorough and constructive comments, which have substantially improved the manuscript. Following their suggestions, we have extensively revised the paper, including major modifications to Sections 2.4, 3.1, 3.2, 3.3 and 4.1. The main changes include (i) a revised evaluation of model performance using complementary statistical metrics (Taylor diagrams, RMSE, Bias, NSE and KGE), (ii) the explicit consideration of reconstruction uncertainties through Monte Carlo analyses, (iii) a clearer and more robust interpretation of global and regional model–data comparisons, and (iv) a substantial reorganization and simplification of the text to improve clarity and readability. In particular, the evaluation of model simulations at regional scale are only focus on the Indian Ocean. Overall, the revised manuscript includes **5 new figures/tables and 12 revised figures/tables** in the main text and Appendix, together with **6 new and 2 revised Supplementary figures/tables**. We believe that these extensive revisions fully address the comments raised by both Reviewer 1 and Reviewer 2 and have significantly strengthened the manuscript.

Reviewer n°1

This manuscript by Barathieu et al moves paleoclimate data-model comparisons in a novel direction, by providing the first example of looking assessing ocean density changes. This is built off the foram-based compilation by Caley et al for the last glacial maximum, combined with the PMIP coupled model simulations.

I believe that this manuscript is a good fit for Climate of the Past and should be published after revision. While I outline some specific comments below, they all come under a single wide umbrella of a comment. I found this article contained so much detail, especially about the individual comparison results) that it obscured your message as times. I feel that you will gain more impact and traction from the manuscript if you are able to be more concise.

Specific comments (not in order of importance).

We are grateful for the positive and constructive comments provided by the reviewer Chris Brierley that helped to significantly improve our manuscript. Below, we present a point-by-point response to the individual comments raised. The reviewer's comments are in black, and our responses are **in purple**.

First, we make an effort to reduce the level of detail in the manuscript in order to improve clarity, although the diversity of model simulations naturally leads to a large amount of material to discuss. We reduced the section 4 as suggested by reviewer 1, concerning regional evaluation, by focusing on the Indian Ocean only.

Sect 2.1.1 You describe the methodology of the Caley et al (2025) data compilation. However, you do not describe any of the broad findings from this dataset. Neither do you provide any information about the uncertainty in the reconstructions, such as an approximate calibration error.

We thank the reviewer for this comment. Since the initial submission, Caley et al. has now been published, and we therefore refer to it as Caley et al. (2026) in the revised manuscript.

We added sentences in section 2.1.1 Surface ocean density to describe the broad findings of the dataset and the uncertainty in the reconstructions. We added after line:

New and published $\delta^{18}\text{O}_c$ datasets were compiled to create an extended database of 474 density reconstructions distributed across all oceanic regions. For each marine sediment core, reconstructions are available for both the LGM and the Late Holocene (LH) (Caley et al., 2026). **The Bayesian hierarchical regression model calibrated to annual surface density yields prediction uncertainties (σ) that vary across species, ranging from 0.48 kg.m⁻³ (*N. pachyderma sinistral*) to 0.86 kg.m⁻³ (*G. bulloides*). More specifically, the mean calibration uncertainties are 0.74 kg.m⁻³ for *G. ruber*, 0.73 kg.m⁻³ for *T. sacculifer*, 0.86 kg.m⁻³ for *G. bulloides*, 0.58 kg.m⁻³ for *N. pachyderma dextral*, and 0.48 kg.m⁻³ for *N. pachyderma sinistral*. When the Bayesian regression model is applied to LGM and LH $\delta^{18}\text{O}_c$ foraminifera databases to reconstruct annual surface density during these periods, we observe stronger increase in LGM surface density value changes at low latitudes compared to mid latitudes.** Analyses from the northern region > 40°N of the Atlantic Ocean were rejected due to potential errors when applying the calibration to the LGM time period (Caley et al., 2026). We thus also exclude this region for the model-data comparison. Surface density is expressed in kg/m³. Throughout this work, values are expressed as anomalies relative to 1000 kg/m³.

L210. I do not see why you would want to interpolate all these results onto a common grid resolution prior to computing the density. Surely it would be more accurate, and computationally efficient to perform all the analysis on each model's native grid.

We interpolated all model outputs to a common 1°×1° grid prior to computing surface density to facilitate direct comparison across model simulations and with the proxy database (also gridded on a 1°×1° grid). A uniform grid simplifies multi-model diagnostics and is widely used in the literature for model-data comparison, particularly when synthesizing results from model simulations with very different native resolutions.

Although computing surface density on each model's native grid could in principle reduce interpolation errors, we expect this does not make a large difference for the broad spatial patterns analyzed. In order to demonstrate this, we calculate the difference between the two approaches with the GISS-E2-R_lgm simulation as an example (see Figure A).

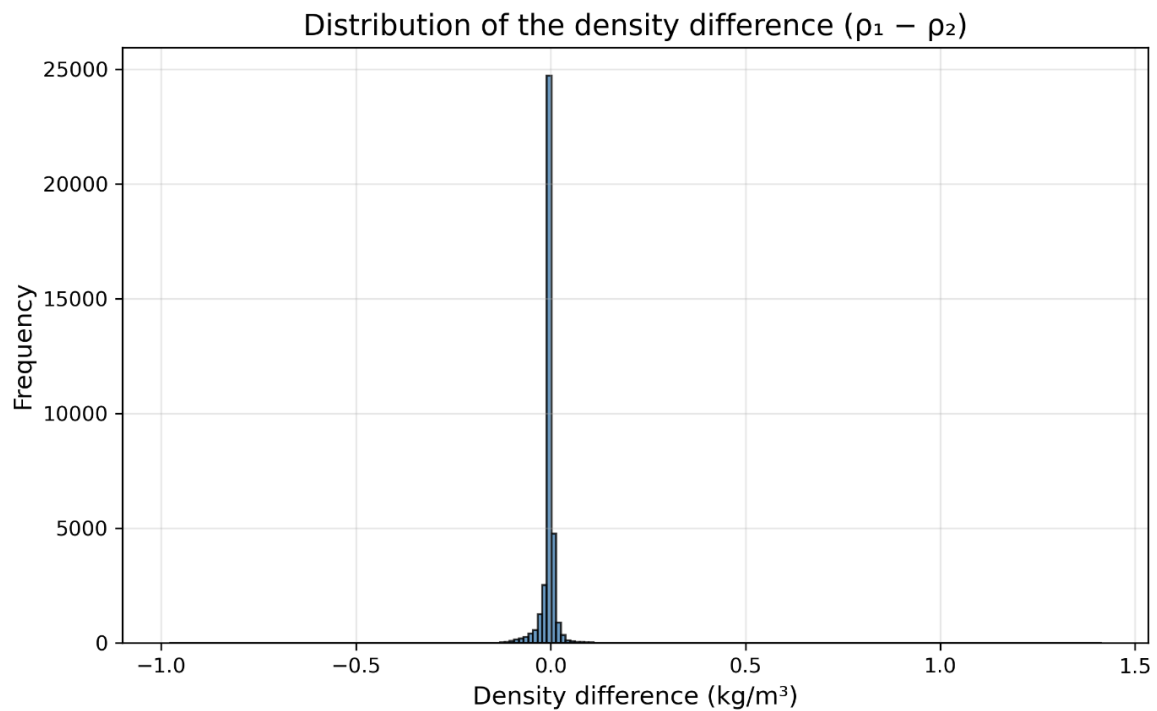


Figure A: Histogram example of the surface density difference between the two approaches for GISS-E2-R_lgm simulation (ρ_1 : density computed with the original resolution and then regridded at $1^\circ \times 1^\circ$, ρ_2 : density recomputed from regridded SST/SSS at $1^\circ \times 1^\circ$). The distribution is strongly peaked around zero, indicating excellent agreement in the open ocean. Slight deviations occur only near coastlines and semi-enclosed seas.

Global statistics indicate that the two density fields are highly consistent over the open ocean. The mean bias is very small (-0.004 kg.m^{-3}), showing that neither method systematically overestimates or underestimates density. The RMSE is also low (0.025 kg.m^{-3}), meaning that typical differences between the two fields are minor.

L214. Is there a reason that you choose to not use any density fields that had be stored on the ESGF?

Not all model simulations provide precomputed surface density fields on the ESGF (iLOVECLIM, MPI-ESM1-2-LR and MIROC-ES2L for example). To ensure consistency across all model simulations, we therefore calculated surface density ourselves from the available temperature and salinity fields.

Starting to work with SST and SSS fields rather than density also allow to check for any potential bias in SST or SSS fields and for the inclusion (or not) of the +1 g/kg of salinity at the LGM, something that could be unclear in the direct density field. Indeed, in two simulations (MIROC-ESM and IPSL-CM5A2), the salinity field was not initialized with the +1 psu offset prescribed in the protocol to account for freshwater stored in ice sheets. To ensure comparability across model simulations, we added +1 psu to the LGM salinity of these two simulations before calculating absolute density.

Table 1. You classify HadCM3 and iLOVECLIM as CMIP6 models. HadCM3 was originally built as part of CMIP3. I accept that it has performed the PMIP4 protocol and is part of PMIP4, but surely not “HadCM3-PMIP3”. Please be more accurate in your model descriptions, as this has implications for your conclusions about the improvement of models between generations (such as in Fig 1).

We thank the reviewer for this important clarification. We agree that HadCM3 itself is not a CMIP6 model and should not be labelled as such. In our study, the PMIP3 designation refers primarily to the ice-sheet boundary condition used in the experiment, not to the model lineage. However, we acknowledge that differences between PMIP phases may also involve additional boundary conditions and forcings (e.g., greenhouse gases, orbital parameters, or other protocol-specific implementations), which may vary across modelling groups. We have revised Table 1 (see below: Revised Table 1) to make this distinction explicit and to avoid any confusion between model generation and boundary conditions. Our comparison is intended to assess the climate response of HadCM3B-M2.1aD under different LGM ice-sheet reconstructions, rather than to imply that the model itself belongs to the PMIP3 generation. Accordingly, interpretations are focused on the role of ice-sheet differences, while acknowledging that other protocol-related differences may also contribute.

L239-241. Please explain what these numbers mean, and why I need to know them.

N were the number of reconstructions by basins. The regional analyses were changed in the revised version so we have added a clarification for the global scale only, in the revised manuscript, L.253:

“At the global scale, the KS threshold is 0.13, consistent with the sample size of our dataset. A KS statistic below the threshold indicates that the model and reconstruction distributions are statistically consistent, whereas values exceeding the threshold indicate a significant difference between distributions.”

L260. Please number the first figure you introduce in your manuscript as Fig 1.

The maps (Fig. 1) have been moved to appear first, and the pseudo-proxy figure is now labelled Fig. 2 in the revised version.

L261. Please describe what the pseudo-proxy approach means here. I have heard the term as a way of combining changes in SST and SSS to give changes in (coral) d18O. But this clearly is not what you mean.

We thank the reviewer for this comment. We clarify that the term “pseudo-proxy” is not used here in the sense of proxy system modelling or the transformation of SST and SSS into a geochemical signal. We have added a clarification in the revised manuscript in section 2.4:

“Here, the pseudo-proxy approach does not refer to proxy system modelling or to the conversion of SST and SSS into a geochemical proxy signal. Instead, it is used as a spatial representativeness test: model values are sampled at the exact locations of the proxy

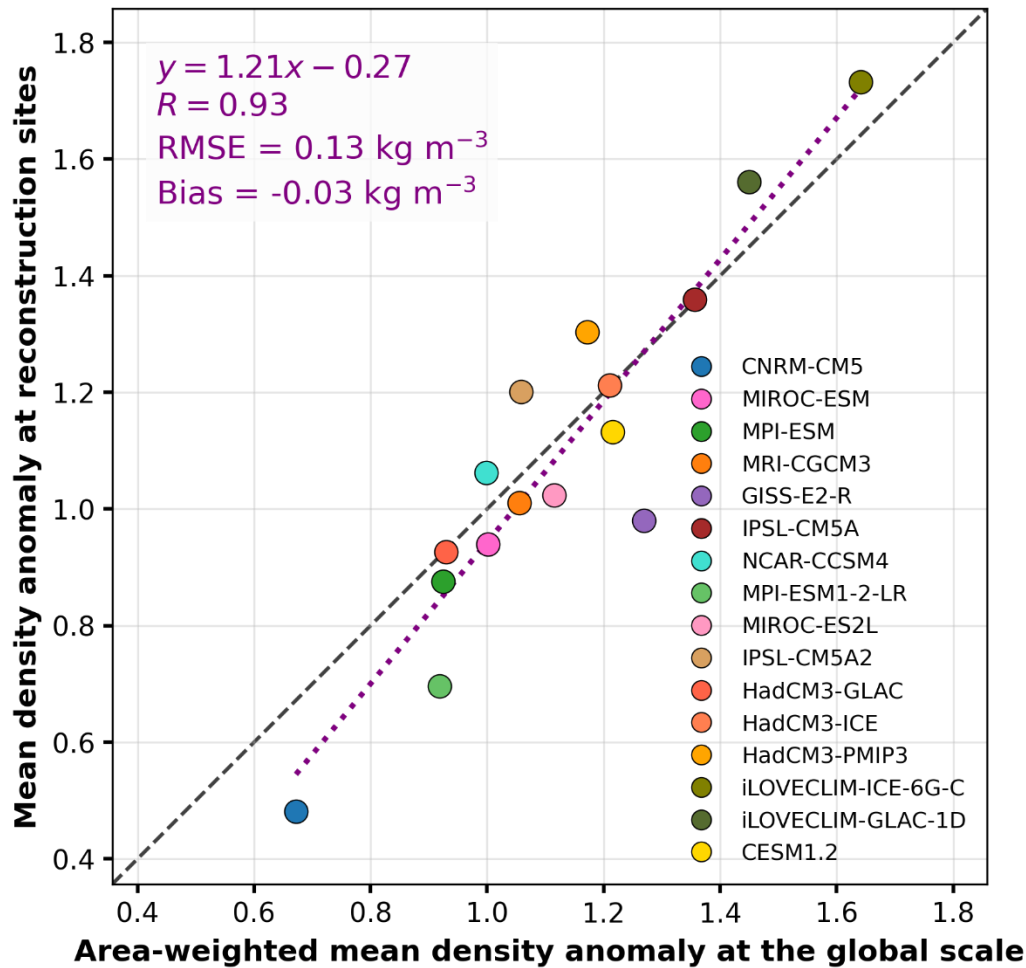
reconstructions and compared with the basin-wide model mean. This allows us to assess whether the uneven proxy network provides an unbiased estimate of the large-scale basin signal.”

Fig. 1. Is this not global? So why is the x-axis labelled ‘in the basin’?

Actually, there was a mistake in how this x-axis was labelled; it is indeed the global axis, and this has now been corrected. Thank you to have bringing this to our attention.

Fig. 1. Consider whether it is appropriate to subdivide between the two model generations.

We agree that subdividing Fig. 1 (now Fig. 2) by model simulation generation is not essential for the message of the figure. Since the remainder of the manuscript does not rely on a PMIP3/PMIP4 distinction, and because each generation is represented by a non-exhaustive ensemble of simulations, we have revised Fig. 1 (now Fig. 2) to show the combined set of simulations rather than splitting them by generation. This makes the figure clearer and avoids suggesting a stronger distinction between generations.

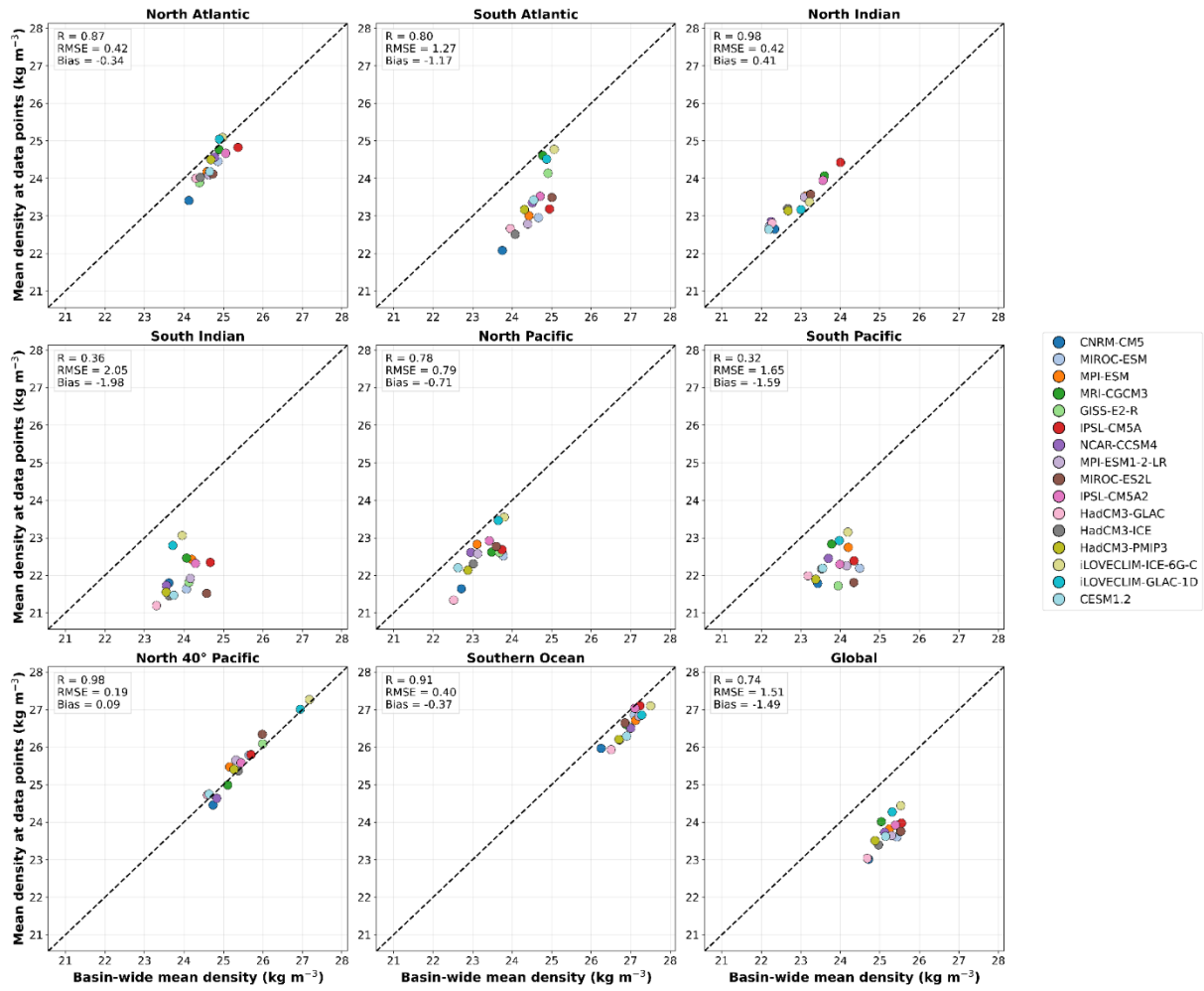


Revised Figure 2: Pseudo-proxy test for global ocean surface density. Comparison between the mean density anomalies (kg/m^3) averaged across all model simulations grid points (x-axis) and the mean density anomalies (kg/m^3) averaged over proxy reconstruction sites (y-axis). The purple regression line and R represent the fit across all model simulations. Results for individual basins are provided in Supplementary Fig.S2.

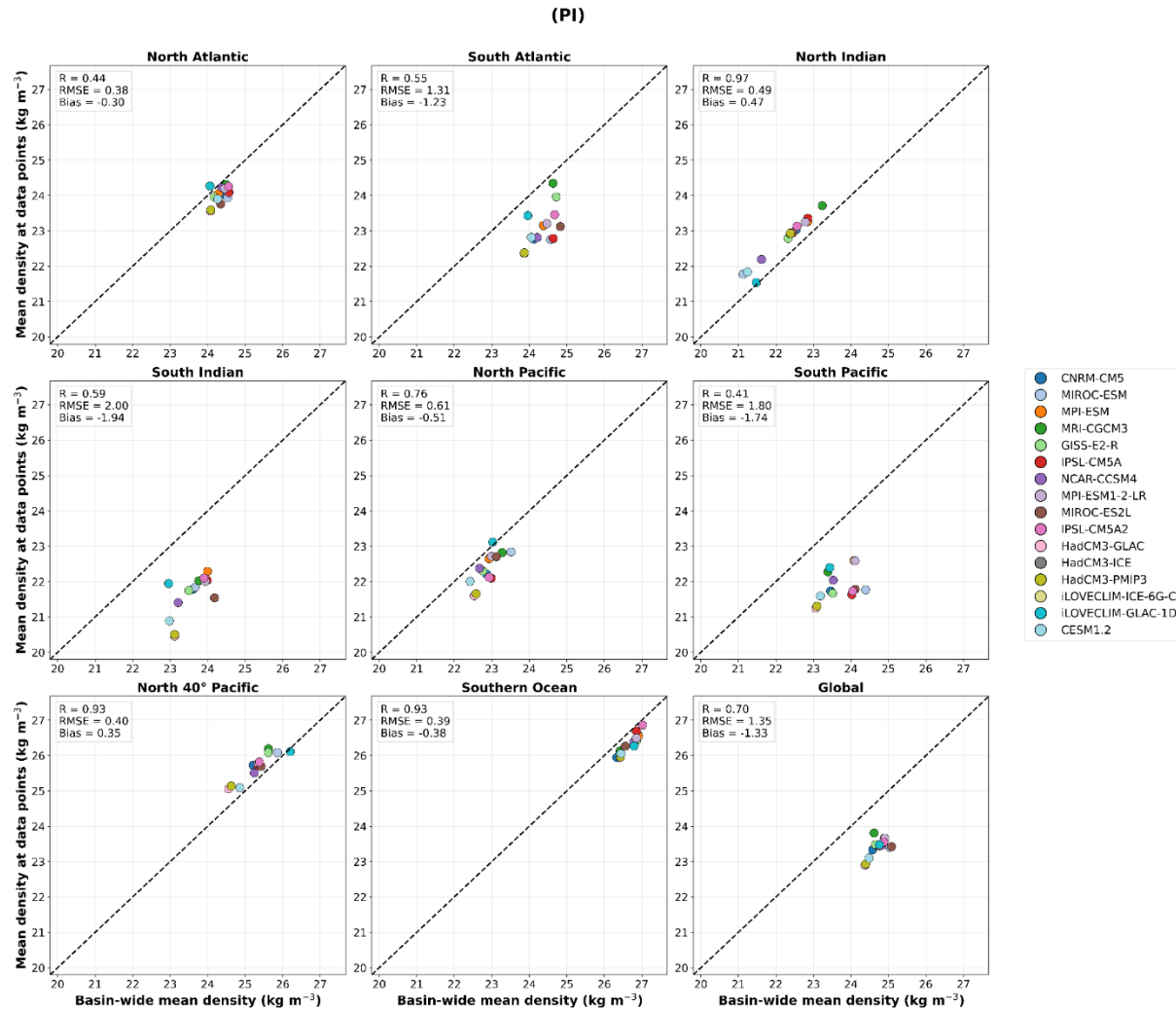
In the Supplementary Material, the basin-scale pseudo-proxy analysis has also been revised by separating the LGM and PI periods. This change ensures consistency with the remainder of the manuscript, where regional density analyses are performed independently for each period rather than using LGM–PI anomalies.

a)

(LGM)



b)



Revised figure S2: Pseudo-proxy test of the spatial representativeness of the proxy network for the LGM (a) and PI (b) ocean surface density. For each ocean basin and for the global ocean, the mean density simulated by each climate model over all grid cells within the basin (x-axis) is compared with the mean density obtained by sampling the same model only at the locations where proxy reconstructions are available (y-axis). Each point represents one model simulation, and the dashed line indicates the 1:1 relationship. This pseudo-proxy approach does not involve proxy system modelling. Instead, it evaluates whether the spatial distribution of proxy sites provides an unbiased estimate of the basin-scale density signal. Statistics shown in each panel include the Pearson correlation coefficient (R), bias, and root mean square error (RMSE) between basin-wide model means and model values sampled at proxy locations. A close agreement with the 1:1 line indicates that the proxy network accurately captures the large-scale basin mean, whereas systematic departures reflect spatial sampling biases arising from the uneven distribution of proxy records. The “Global” panel includes all proxy locations from the individual basins shown.

(e.g.) L303. Why do you write **absolute** density anomalies. Do relative density anomalies have any meaning (I guess they will all be around 0.1%)? So what is the ‘absolute’ clarifying?

We thank the reviewer for this helpful comment. We agree that the term “absolute” may be unclear and misleading, and we have revised the text to define it more explicitly and to distinguish the full LGM density state from the anomaly associated only with hydrographic SST and SSS changes.

Here is the revised version of the entire paragraph:

“The LGM experimental protocol (Kageyama et al., 2017) considered as boundary conditions the large continental ice sheets, the associated land–sea mask changes, adjustments to ocean salinity (as ice sheets store large volumes of freshwater), and reductions in greenhouse gases. This makes it a challenging experiment for climate models, which explains why only a limited number of modelling groups have performed it. In two simulations (MIROC-ESM and IPSL-CM5A2), the salinity field was not initialized with the +1 psu offset prescribed in the protocol to account for freshwater stored in ice sheets. To ensure comparability across model simulations, we added +1 psu to the LGM salinity of these two simulations before calculating absolute density. **Here, “absolute” surface density refers to the full LGM density state, including the mean ocean density increase associated with reduced ocean volume and the corresponding global salinity offset prescribed by the LGM protocol. It therefore differs from the hydrographic density anomaly driven only by local changes in SST and SSS. To ensure comparability across simulations, we first place all model outputs on the same absolute-density baseline by applying the prescribed +1 psu correction where needed.** Simulations performed with the iLOVECLIM model found the dynamical effect of that +1 psu to be very small, supporting this direct correction (Caley et al., 2026). For the calculation of surface density changes due to the hydrographic changes in SST and SSS, i.e. corrected for mean ocean density changes related to ocean volume, we removed this +1 psu from the LGM simulations and applied a -0.77 kg/m^3 density correction to the reconstructions, following Caley et al. (2026).”

L308. Please put the values in Table 1.

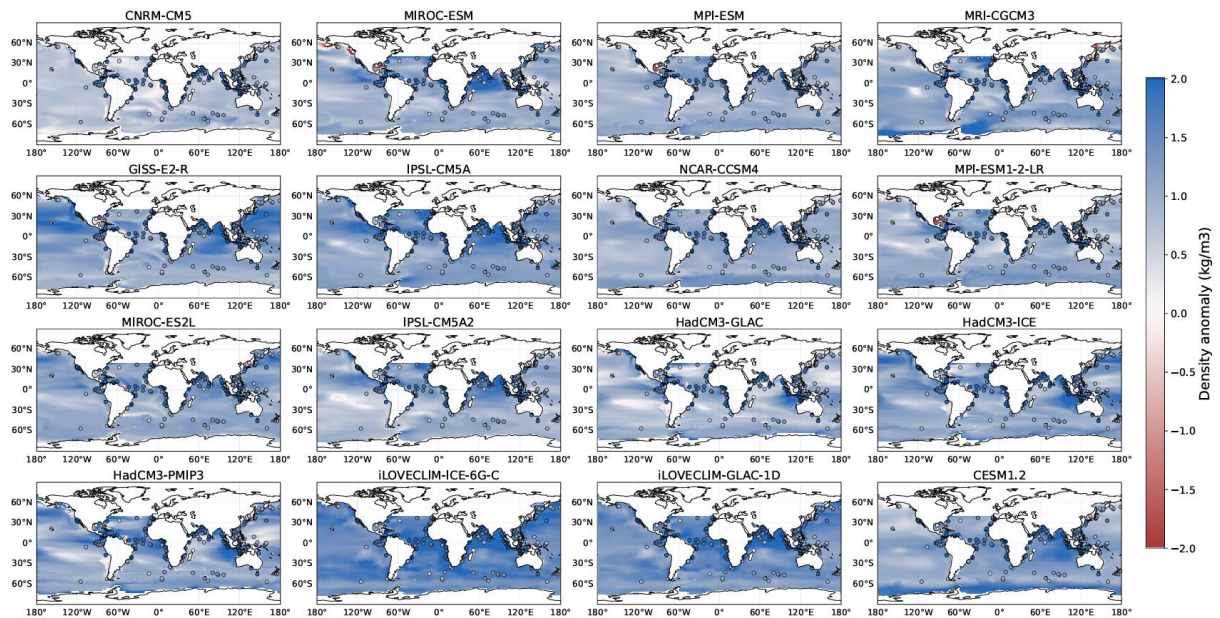
We thank the reviewer for this suggestion. We have computed the mean LGM–PI surface density anomaly for each model simulation, using a latitude-weighted spatial average, and we now include these values in Table 1, as individual model simulations estimates, and as an ensemble mean over all simulations in the text.

Model simulation	Figure label	PMIP/CMIP phase and boundary conditions	Mean LGM-PI surf. density anomaly (kg m^{-3})	References
CNRM-CM5	CNRM-CM5	PMIP3-CMIP5 r1ilp1	0.672	Sénési et al. (2014 - LGM and PI) Voldoire et al. (2013)
MIROC-ESM	MIROC-ESM	PMIP3-CMIP5 r1ilp1	1.002	JAMSTEC et al. (2015 - LGM and PI) Sueyoshi et al. (2013)
MPI-ESM-P	MPI-ESM-P	PMIP3-CMIP5 r1ilp1	0.924	Jungclaus et al. (2014 - LGM and PI) Adloff et al. (2018)
MRI-CGCM3	MRI-CGCM3	PMIP3-CMIP5 r1ilp1	1.055	Yukimoto et al. (2012) Yukimoto et al. (2015 - LGM and PI)
GISS-E2-R	GISS-E2-R	PMIP3-CMIP5 r1ilp150	1.269	NASA/GISS (2014 - LGM and PI) Schmidt et al. (2014) Ullman et al. (2014)
IPSL-CM5A-LR	IPSL-CM5A	PMIP3-CMIP5 r1ilp1	1.357	Kageyama et al. (2016 - LGM) Caubel et al. (2016 - PI)
CCSM4	NCAR-CCSM4	PMIP3-CMIP5 r1ilp1	0.998	Gent et al. (2011) Brady et al. (2013)
<i>HadCM3B-M2.1aD</i>	<i>HadCM3-PMIP3</i>	<i>PMIP4-CMIP6</i> <i>PMIP3 ice sheets</i>	1.172	<i>Izumi et al. (2023)</i>
<i>HadCM3B-M2.1aD</i>	<i>HadCM3-GLAC</i>	<i>PMIP4-CMIP6</i> <i>GLAC-1D ice sheet</i>	0.929	<i>Izumi et al. (2023)</i>
<i>HadCM3B-M2.1aD</i>	<i>HadCM3-ICE</i>	<i>PMIP4-CMIP6</i> <i>ICE-6G-C ice sheet</i>	1.211	<i>Izumi et al. (2023)</i>
<i>MPI-ESM1-2-LR</i>	<i>MPI-ESM1-2-LR</i>	<i>PMIP4-CMIP6</i> <i>r1ilp1f1</i>	0.919	<i>Mauritsen et al. (2019)</i>
<i>MIROC-ES2L</i>	<i>MIROC-ES2L</i>	<i>PMIP4-CMIP6</i> <i>r1ilp1f2</i>	1.115	<i>Ohgaito et al. (2019 - LGM)</i> <i>Hajima et al. (2020, 2019 - PI)</i>
<i>IPSL-CM5A2</i>	<i>IPSL-CM5A2</i>	<i>PMIP4-CMIP6</i>	1.058	<i>Sepulchre et al. (2020)</i>
<i>iLOVECLIM1-1-1-GLAC-1D</i>	<i>iLOVECLIM-GLAC-1D</i>	<i>PMIP4-CMIP6</i> <i>GLAC-1D ice sheet</i>	1.450	<i>Lhardy et al. (2021)</i> <i>Bouttes et al. (2023)</i>
<i>iLOVECLIM1-1-1-ICE-6G-C</i>	<i>iLOVECLIM-ICE-6G-C</i>	<i>PMIP4-CMIP6</i> <i>ICE-6G-C ice sheet</i>	1.641	<i>Lhardy et al. (2021)</i> <i>Bouttes et al. (2023)</i>
<i>CESM1-2</i>	<i>CESM1.2</i>	<i>PMIP4-CMIP6</i>	1.215	<i>Hurrell et al. (2013)</i> <i>Tierney et al. (2020)</i>

Revised Table 1: Model simulations available for this study, with LGM and piControl simulations. Model simulations from PMIP3 and PMIP4 (in italics) and mean LGM-PI surface density anomaly for each simulation (kg/m^3). References correspond to the model description papers and/or ESGF simulation references when available.

Fig. 2 This is very hard to read, and especially see the difference between the model and proxy data. Consider things like different projections to minimise the amount of white space. Maybe consider a separate ensemble-mean map, which could be larger? Provide the units of the anomaly

We agree that Fig. 2 (now called Fig. 1) is visually dense and that the differences between model simulations and proxy data are hard to read. We have reduced the white space between the figures and provided the unit of the anomaly (see revised Fig. 1). Given the number of simulations we find it difficult to improve the readability. However, we have added a new figure in the supplementary material showing the ensemble-mean density anomaly over all model simulations. This makes it easier to see where the proxy sites are.



Revised Figure 1: Absolute surface density (kg/m^3) anomaly map (LGM - piControl). The dots represent the surface density anomaly database reconstructions (Caley et al., 2026) and the background map is the anomaly mean for each model simulations in the study.

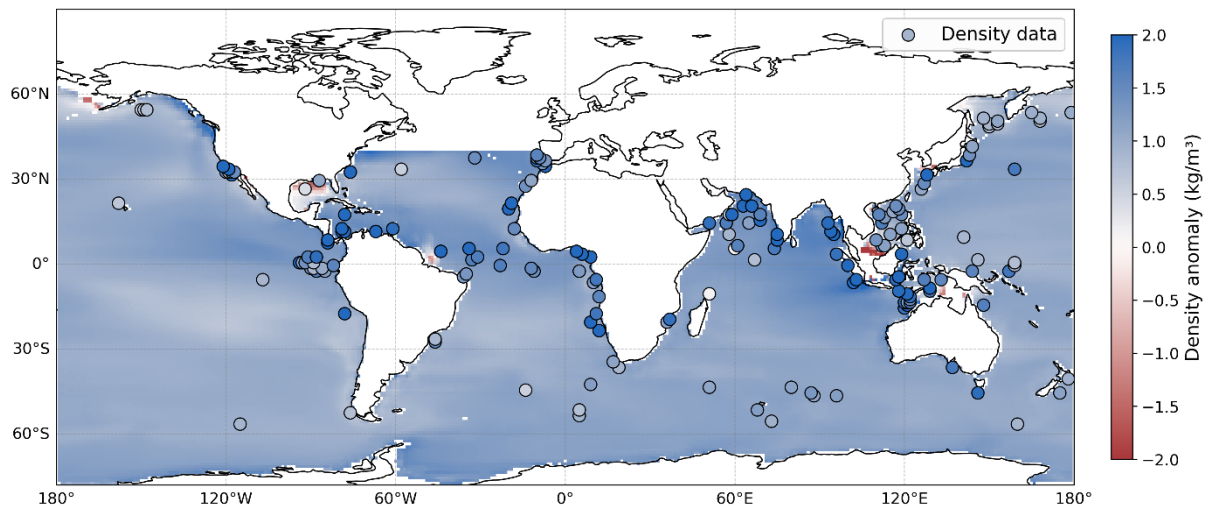
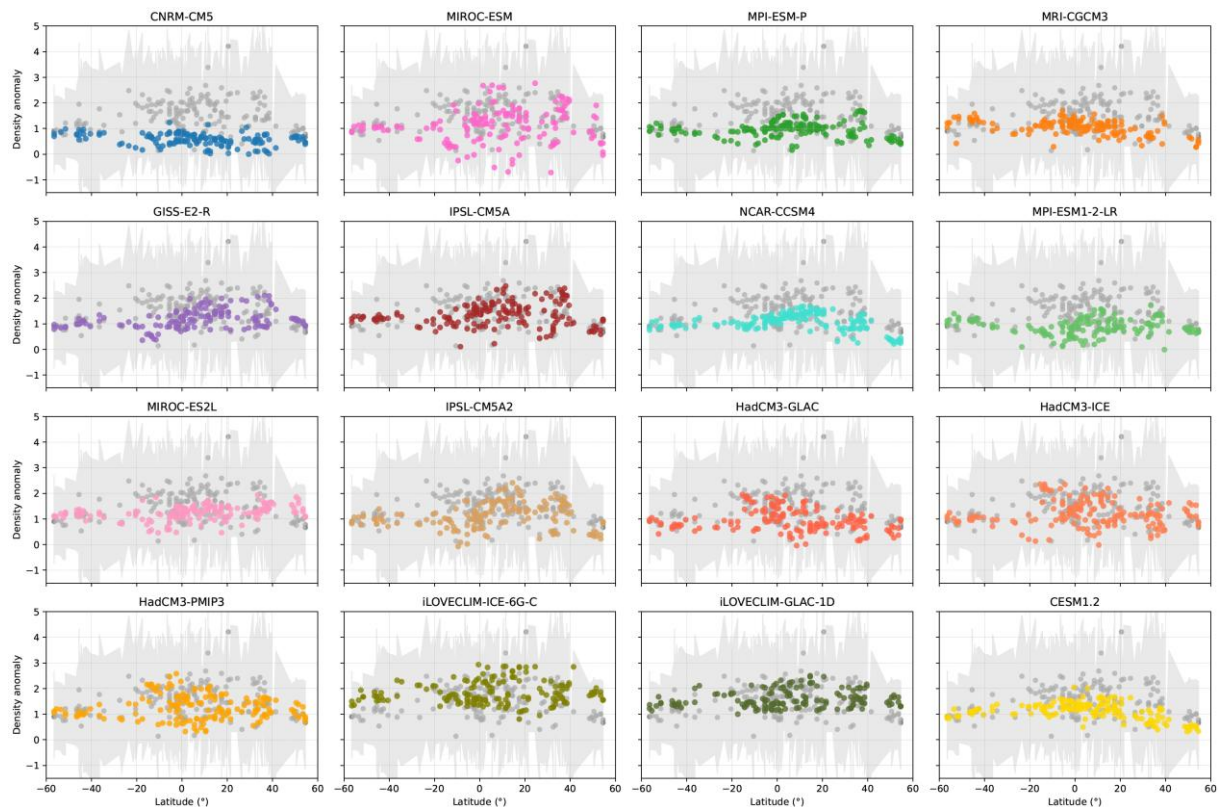


Figure S1 (new): Absolute surface density (kg/m^3) anomaly map (LGM - piControl). The dots represent the surface density anomaly database reconstructions (Caley et al., 2026) and the background map is the anomaly mean for all model simulations in the study.

Fig. 3. I find this very hard to interpret. Why not remove the individual data points and plot as zonal means.

We agree that Fig. 3 is visually dense, and we have indeed experimented with zonal mean plots alone. However, we chose to keep the individual proxy points in order to preserve information about the spatial variability and scatter of the observations, which can otherwise be masked when only zonal means are shown. The current version allows the reader to see both the general latitudinal trend and the spread of the data.



Revised Figure 3: Density anomaly (kg/m^3) as a function of latitude for each model simulation (colored dots) compared with the observational data (grey dots) and the 95% confidence interval (grey shading). Model outliers, identified using the interquartile range (IQR) method (values outside $1.5 \times \text{IQR}$), were excluded to reduce the influence of extreme values. This filtering highlights the main structure and latitudinal patterns of the modelled density anomalies while retaining all available latitude points.

L348. Presumably you have been undertaking this decomposition at individual grid points, and comparing to the uncertainty at in each proxy reconstruction. However, I can't see how this results in the Fig. 4. The bars look like there are guaranteed to add to 100%, but is there a bit of non-linearity in the equation of state. How have you accounted for this?

Fig. 4: Can you provide an estimate of the uncertainty in your decomposition?

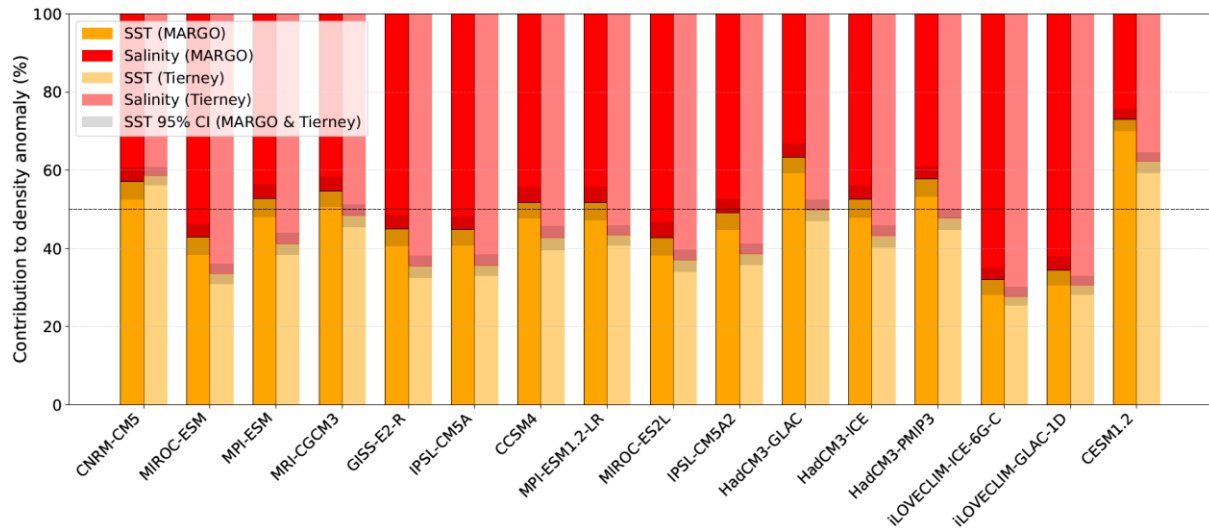
We thank the reviewer for these important comments.

In the initial version of Fig. 4, the contributions of SST and salinity were computed sequentially (temperature first, then salinity), which implicitly assumes a linear decomposition. As the reviewer correctly points out, the equation of state is non-linear, and therefore the decomposition depends on the order in which temperature and salinity anomalies are applied.

To address this, we implemented a Shapley decomposition, which provides an order-independent attribution by averaging the contributions obtained from the two possible sequences (temperature-first and salinity-first). This approach ensures that the total density anomaly is exactly partitioned while accounting for non-linear interactions.

We also explicitly quantified the magnitude of the non-linear term as half the difference between the two decompositions. We find that this contribution is negligible (on the order of $\sim 10^{-4} \text{ kg/m}^3$), indicating that non-linear effects do not significantly affect our conclusions.

Finally, to estimate uncertainties, we performed a bootstrap analysis by sampling SST anomalies within their proxy uncertainty ranges. This allows us to derive 95% confidence intervals on the SST and salinity contributions shown in Fig. 4.



Revised Figure 4: Relative contributions (%) of SST (orange) and SSS (red) to the model–data density anomaly (LGM–PI) differences. Contributions are expressed as percentages of the total absolute density difference between model simulations and reconstructions. Darker bars correspond to the MARGO SST dataset (MARGO Project, 2009) while lighter bars correspond to the Tierney et al. (2020) SST dataset. The 95 % confidence interval (CI) of the SST contribution is shown as a grey band (CI) overlapping each bar, obtained from a bootstrap over the observational SST uncertainty and the Shapley decomposition. The decomposition first computes the SST-only density anomaly (using model simulation LGM temperatures with PI salinity) and then attributes the residual to salinity effects, taking into account the non-linearity of the equation of state via TEOS-10 GSW routines.

L380. Remove repeated word: reconstruct

Done.

We revised for: “Model simulations that best reproduce the density anomalies inferred from proxy data also tend to exhibit the largest reductions in tropical precipitation”.

Fig. 5. Shouldn’t these regression lines go through the origin, by definition of being anomalies? What is the implication of the offsets?

We thank the reviewer for this insightful comment. The confusion arises from our use of the term “anomaly”. In this study, “anomalies” are defined as differences between two climatic states (LGM minus PI), rather than deviations from a mean state. As such, they are not expected to be centred around zero, and the regression line is therefore not constrained to pass through the origin.

The non-zero intercept reflects residual differences in the mean LGM–PI response between precipitation and density across model simulations.

We have clarified this definition in the revised manuscript to avoid ambiguity, adding this sentence L.246:

“In this study, the term “anomaly” refers to the difference between two climatic states (LGM minus PI), following common usage in paleoclimate studies, and does not imply a deviation from a mean state.”

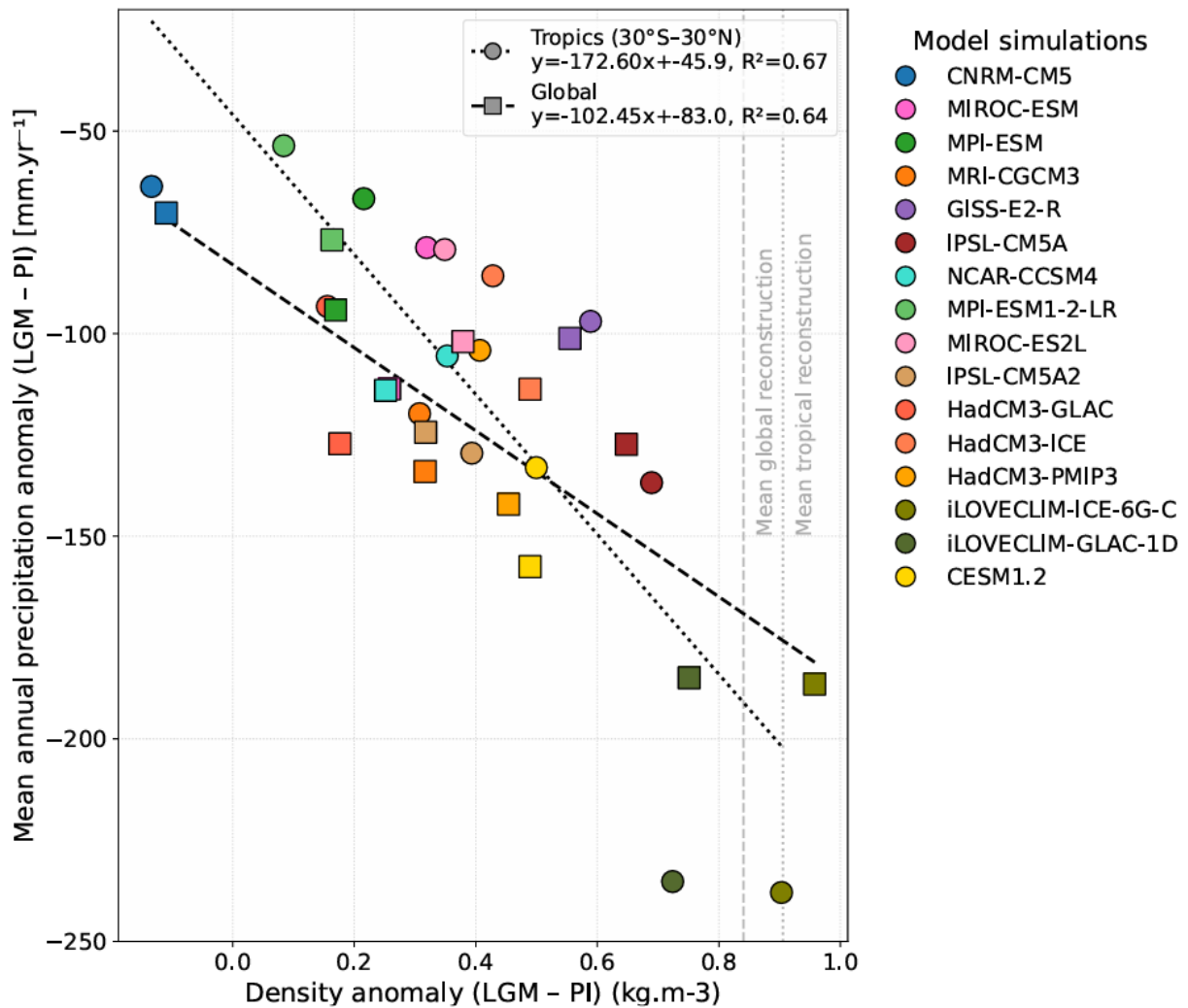
Fig. 5. Previously you stated the global reconstruction mean of the density anomaly was 1.5 (L310), but here it is shown as 0.8. Please clarify!

We thank the reviewer for this careful observation. The value of 1.5 kg/m³ mentioned in the text (L310) corresponded to the mean absolute density anomaly over the global reconstruction domain, while the value of 0.8 kg/m³ shown in Fig. 5 was initially derived from the non-absolute form of the anomaly (i.e., driven only by regional hydrographic changes in SST and SSS, without global ice-sheets effect on global ocean salinity), it can be simply calculated as the absolute anomaly minus a constant offset of 0.77 kg/m³ as explain before and in Caley et al., 2026).

In the revised version, the global mean reconstruction on the x-axis of Fig. 5 is 0.84 kg/m³, while the absolute mean anomaly used in the text is 1.61 kg/m³. Taking into account the offset of 0.77 kg/m³, this is consistent:

$$0.84 \approx 1.61 - 0.77$$

The updated figure therefore reflects the corrected, consistent treatment of the density difference (LGM–PI) used throughout the manuscript.



Revised Figure 5: Mean annual precipitation anomalies (LGM - PI, mm.yr⁻¹, over land and ocean) as a function of seawater density anomalies corrected for mean ocean density changes (kg/m³) relative to sea level-induced salinity increase at LGM. Scatter points show the relationship for the tropics (30°S–30°N, circle) and the global ocean (square). Dashed lines indicate linear regressions for each region, with the corresponding slope, intercept, and R². Precipitation anomalies are from Kageyama et al. (2021). Grey dashed lines indicate the mean tropical and global reconstructions. No linear relationship is found between SST anomalies and mean annual precipitation anomalies (not shown), indicating that the link between density anomalies and precipitation is primarily driven by salinity changes.

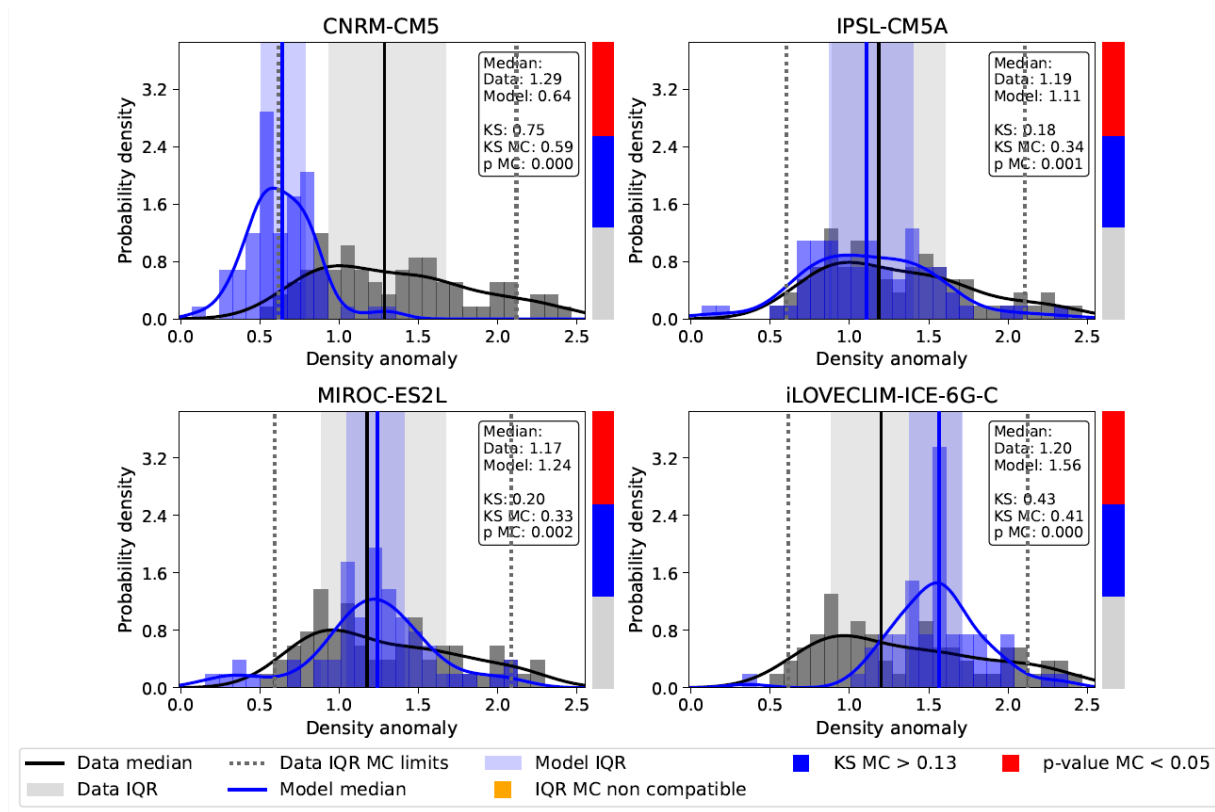
L417. I believe that the proxy reconstruction IQRs are the range of the values average across the globe (or subsequently a region). But surely there is a calibration error attached to the reconstruction. How are you treating this uncertainty, and how does it alter the IQR?

We thank the reviewer for this valuable suggestion, which led us to substantially improve the statistical comparison between model simulations and proxy reconstructions.

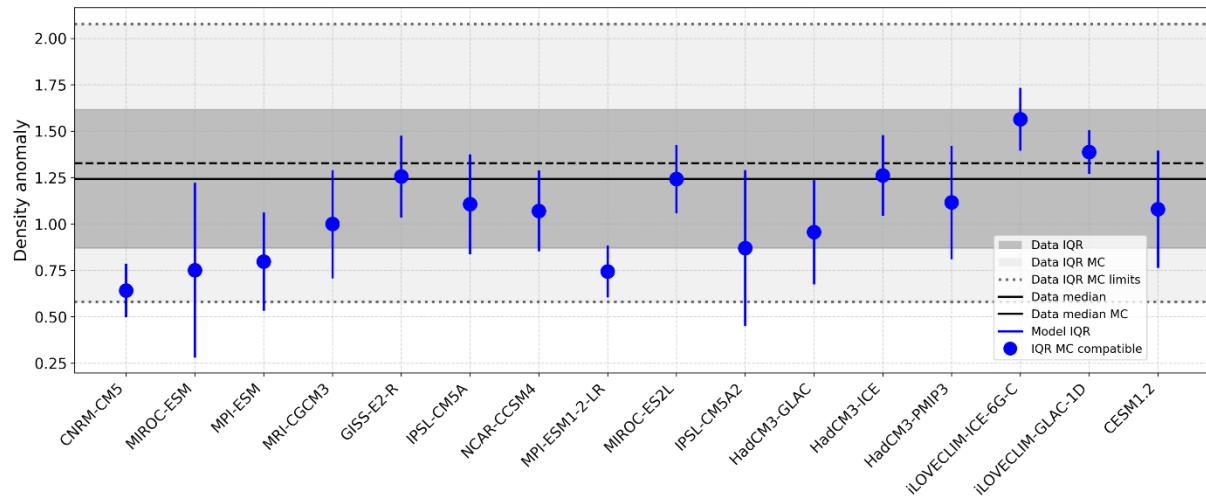
Following this comment, we explicitly propagated the calibration uncertainty associated with each density reconstruction through a Monte Carlo procedure. For each reconstruction site, random realizations were generated within the 95% confidence interval derived from the Bayesian density calibration of Caley et al. (2026). These Monte Carlo realizations were then

used to recompute the reconstructed distributions, their interquartile ranges (IQRs), and the associated statistical diagnostics.

This revision revealed that accounting for calibration uncertainty substantially broadens the reconstructed IQR, as expected. Consequently, the IQR compatibility criterion becomes less discriminating: while two model simulations previously fell outside the reconstructed IQR, all sixteen simulations become compatible with the Monte Carlo-propagated reconstruction envelope once uncertainty is taken into account (new Figs. 6 and 7). However, despite this broader uncertainty range, the Monte Carlo-derived Kolmogorov–Smirnov statistics and associated p-values remain outside the prescribed acceptance thresholds for all simulations. Therefore, our main conclusion is strengthened rather than weakened: accounting for calibration uncertainty reconciles the overall spread of reconstructed and simulated density anomalies, but significant discrepancies remain in the overall shape of their distributions.



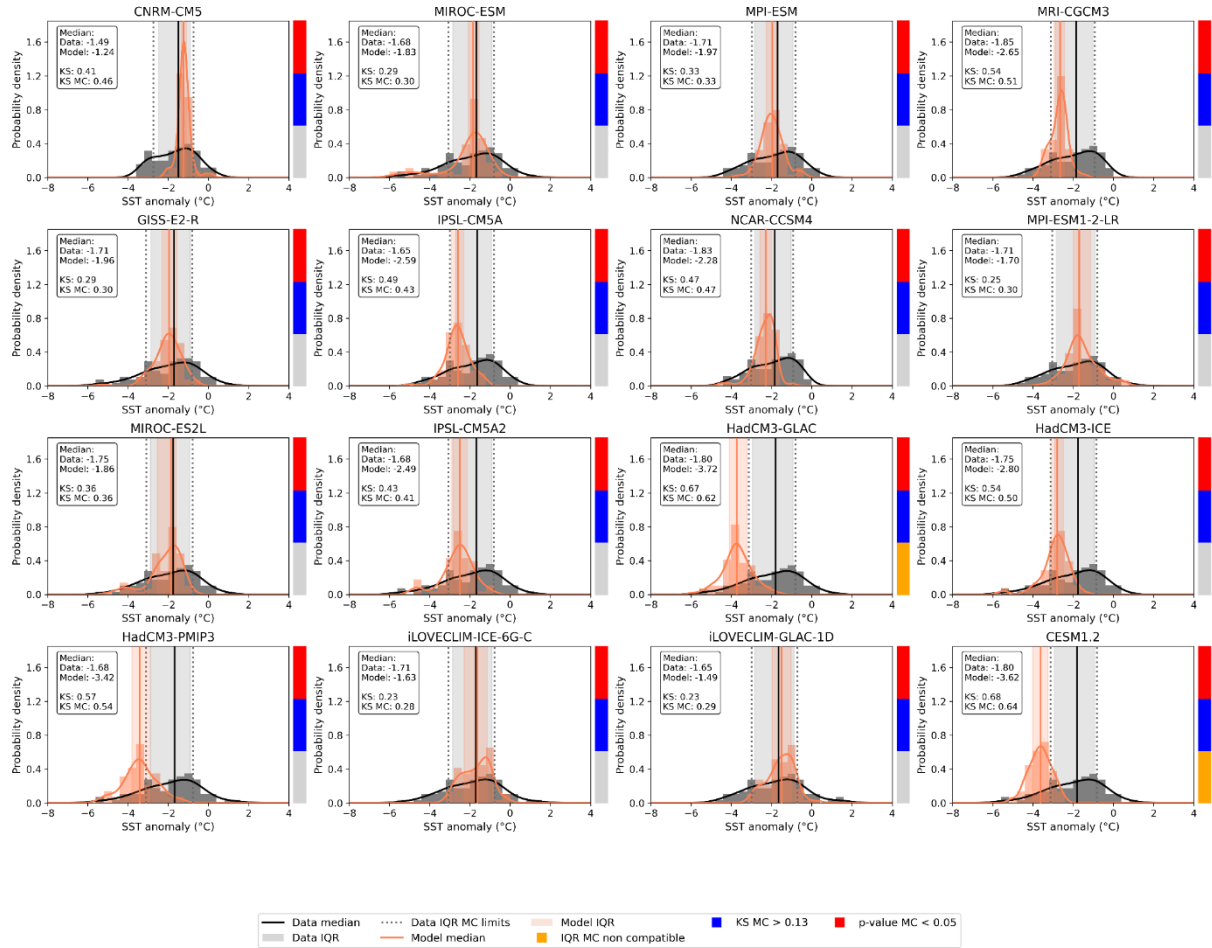
Revised Figure 6: Distribution histograms of surface density anomalies (LGM-PI, kg/m³) for an example of four model simulations. Density reconstructions are shown in black, model simulations in blue. Kernel Density Estimates (KDEs) illustrate the central tendency and overall shape of the distributions. 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.



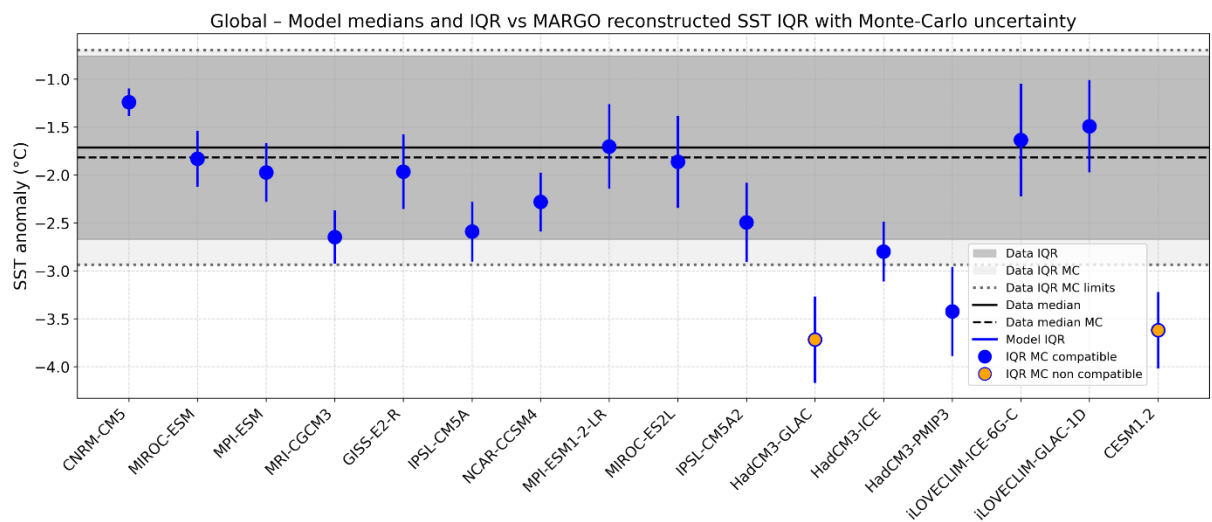
Revised Figure 7: The dark grey band indicates the reconstructed IQR without considering uncertainties, while the light grey band and grey dotted lines indicate the Monte Carlo propagated reconstruction IQR after accounting for reconstruction uncertainty. The solid black line marks the reconstructed 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.

Following the same methodology, we also propagated reconstruction uncertainties for the SST comparisons using both the MARGO (MARGO Project, 2009) and Tierney et al. (2020) datasets (new Supplementary Figs. S3 and S4). As for density, uncertainty propagation increases the reconstructed IQR and leads most model simulations to satisfy the IQR criterion, whereas the Monte Carlo-derived KS statistics and p-values continue to indicate statistically significant differences between reconstructed and simulated distributions.

a)



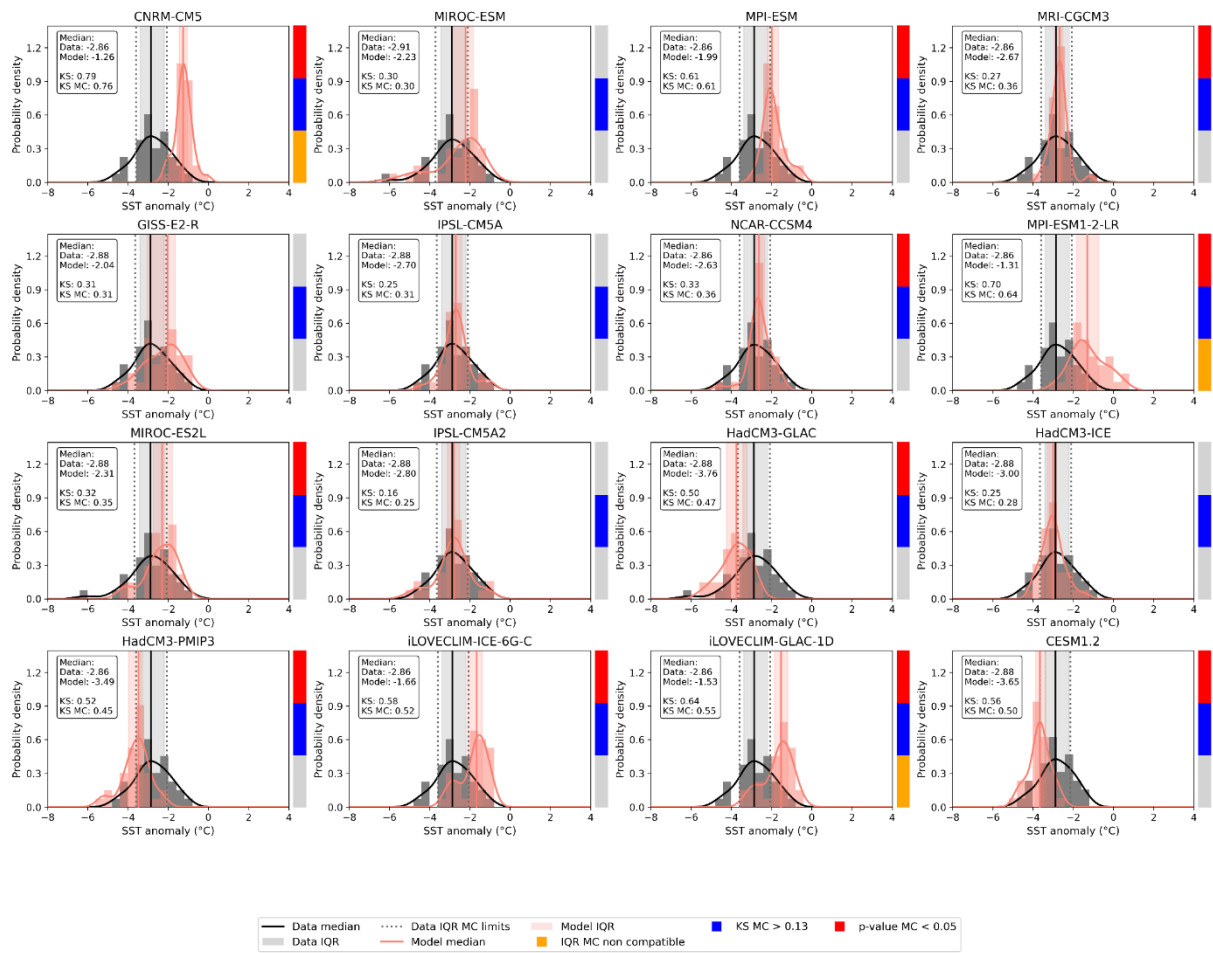
b)



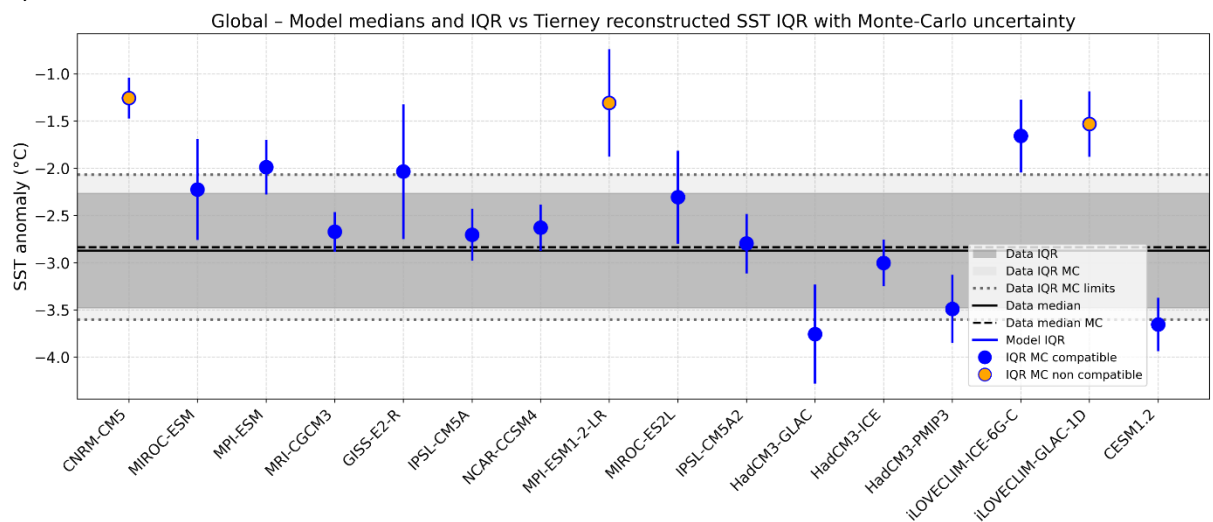
Revised Figure S3: (a) Distribution histograms of SST anomalies (LGM-PI, °C) at the global scale. SST reconstructions from the MARGO database (MARGO project, 2009) are shown in black, and model simulations in orange. Kernel Density Estimates (KDEs) illustrate the central tendency and shape of the distributions. Vertical lines indicate medians, and shaded envelopes represent interquartile ranges (IQRs), capturing spread independent of outliers. Histograms show frequency and complement the KDEs. Extreme values were excluded via IQR filtering. 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)



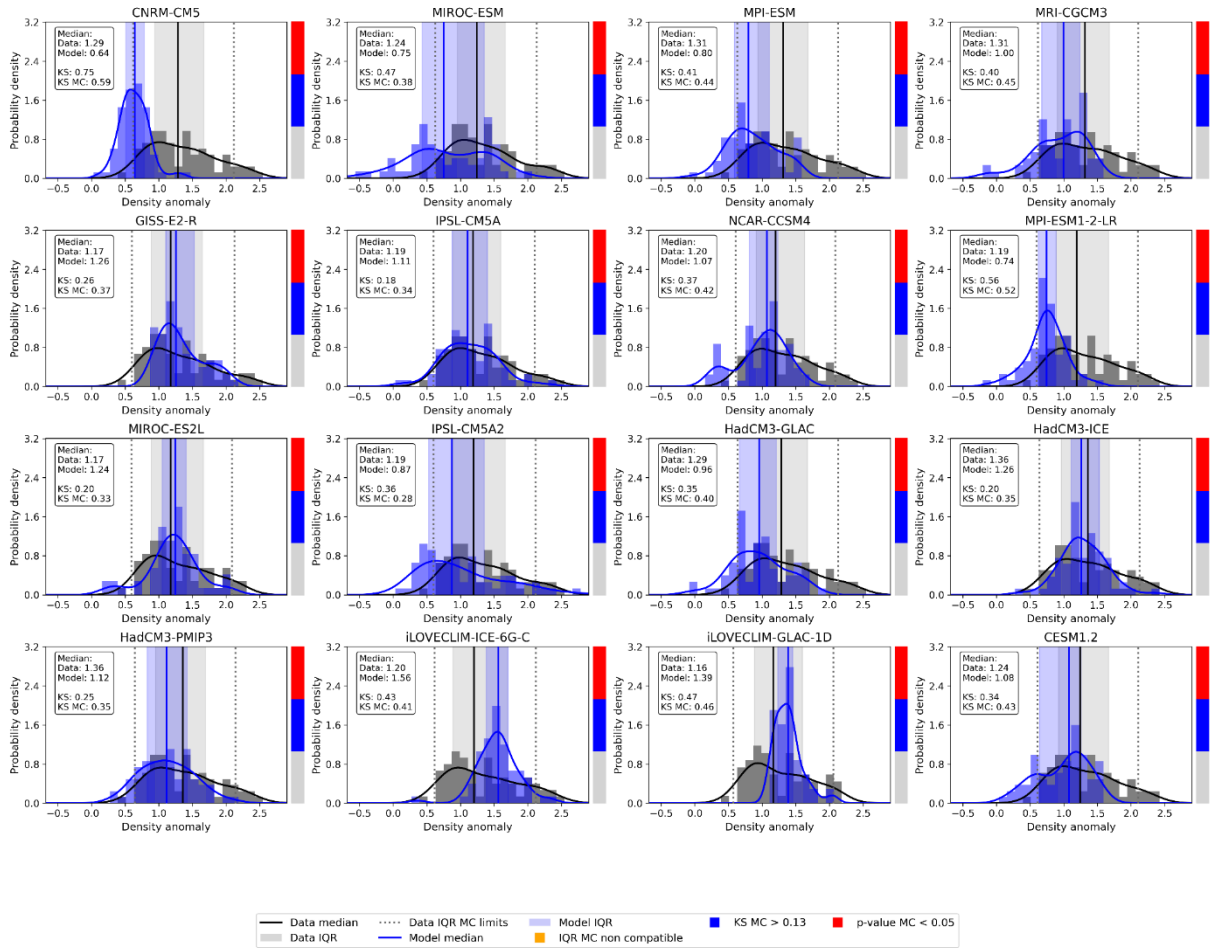
b)



Revised Figure S4: (a) Distribution histograms of SST anomalies (LGM-PI, °C) at the global scale. 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. Extreme values were excluded via IQR filtering. 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.

The manuscript has been extensively revised to describe this new methodology (Sect. 2.2), present the updated figures (Figs. 6 and 7 and Supplementary Figs. S3 and S4), and discuss its implications in Sect. 3.2.



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.

Fig. 8. I do not understand what this figure is assessing, and how to interpret that from the analysis. Is it trying to measure the ability of the models to capture the spatial pattern of surface density at both the PI and LGM? If so, why don't you present some maps. Or even better move to using a Taylor Diagram, to summarise the observational comparison?

Fig. 8 is not intended to assess the spatial pattern of surface density itself, but rather to evaluate how well each model simulation reproduces the relationship between reconstructed and simulated density values at individual sites, for both the PI (piControl) and LGM states. However, as each point corresponds to a specific geographical location, this analysis provides indeed a useful assessment of how well the model simulation reproduces the spatial pattern of surface density across the proxy sites. For each model simulation, we perform a global regression of simulated density against reconstructed density, and Fig. 8 displays the resulting regression lines, R^2 values, and slopes, with error bars accounting for observational uncertainty via Monte Carlo perturbation of the 95 % confidence intervals. In this context, the R^2 values quantify how well the model simulations captures the spatial pattern (that is inter-site differences) observed in the reconstructions, while the slope provides information on the amplitude of the response. The dashed 1:1 line allows a visual comparison of model simulations bias and variance against the observations.

In the revised version, as suggested by the reviewer, we have created Taylor diagrams (below) and we also included additional statistical parameters as suggested by Reviewer 2 (see response to Reviewer 2).

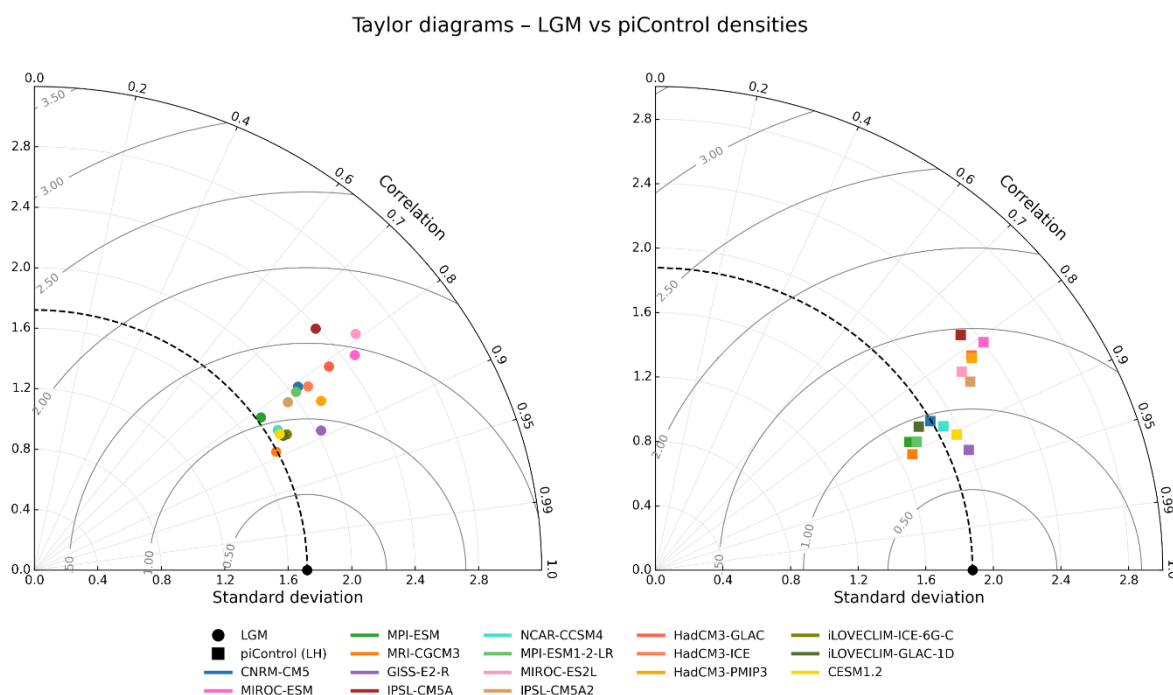


Figure B: Taylor diagrams comparing model simulations performance for LGM (left) and piControl (LH, right) ocean surface density reconstructions. All grid cells are equally weighted in these diagrams. Each point represents a climate model simulation. Circles correspond to the Last Glacial Maximum (LGM), while squares correspond to the pre-industrial control simulation (piControl, LH). The radial distance indicates the model simulation standard deviation normalized to the reference observations, while the angular coordinate represents the pattern correlation between model outputs and reconstructions. The black dots define the reference observational standard deviation which represents the spatial variability of the observational reconstructions, computed as the standard deviation of all observations within each period. Contours show centered root-mean-square error (CRMSE) relative to observations.

Interpretation of these Taylor diagrams were added L.528 :

“Model simulations performance was evaluated using Taylor diagrams (Fig. 8), which summarize the spatial correlation, normalized standard deviation and centered root-mean-square error (CRMSE) between simulated and reconstructed surface densities, together with complementary statistical metrics (RMSE, bias, NSE and KGE; Fig. 9). To complement these metrics, scatter plots comparing reconstructed and simulated surface densities at each reconstruction site are provided in Fig. D1 (see Fig. S5 for SST). This multi-metric approach provides a more comprehensive assessment of model skill by simultaneously evaluating the spatial pattern, variability, systematic bias and overall agreement with the reconstructions.”

L531. You perform a basin-by-basin analysis. But your previous section concluded the tropics and extratropics behaved quite differently. However, you chosen analysis folds these two regions into the same basin. Please justify your choice.

We thank the reviewer for this important comment, which led us to substantially reorganise the regional analysis.

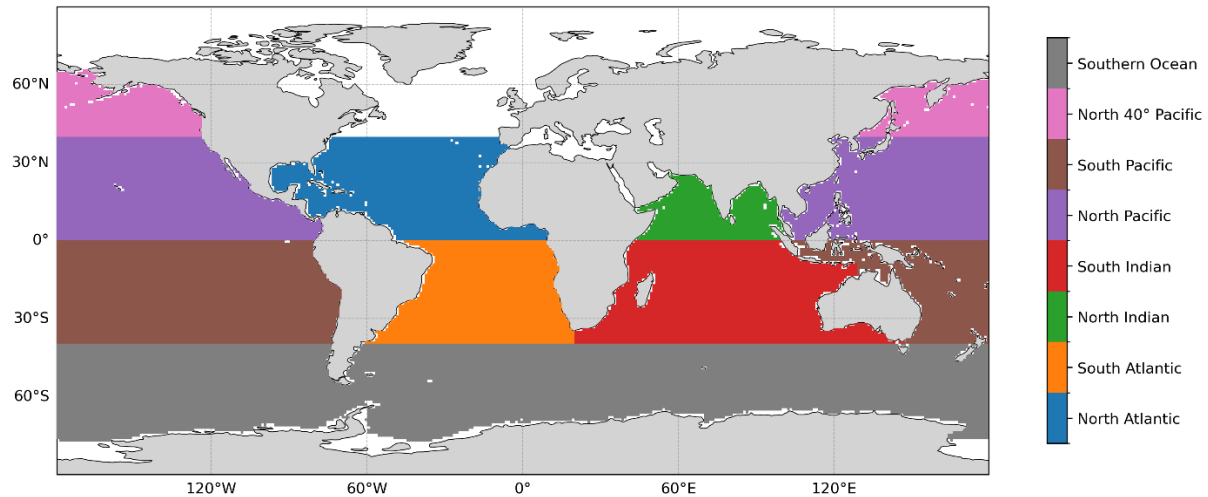
In the revised manuscript, rather than discussing all ocean basins equally, we now focus the main text exclusively on the Indian Ocean. This region is crucial for tropical–extratropical interactions and has a high density of reconstructions. It is a place of major hydroclimate processes such as monsoon and Walker circulation dynamics and previous studies already conducted some data model comparison in this region, enlightening its complexity. The Indian Ocean is explicitly divided into a northern and southern sector, separated at the equator (0°), reflecting the strong dynamical asymmetry between hemispheres. The southern boundary of the basin is set at 40°S, consistent with the present day location of the subtropical convergence, as well as the connection between the Indian and Atlantic Ocean through the greater Agulhas current system.

This new organisation allows us to directly compare model performance between the monsoon-dominated North Indian Ocean and the subtropical South Indian Ocean, while providing a natural transition towards the West–East density gradient analysis presented in the following section.

The analysis of other basins (Atlantic, Pacific and Southern Ocean) has been retained but moved to the Supplementary Material, where we apply consistent latitude-based partitions

(0°, 40°N, 40°S) to ensure comparability across basins and separate tropical from extratropical circulation dynamics (Revised Fig. A1).

Overall, this revision directly addresses the reviewer's concern by explicitly distinguishing tropical and subtropical regimes within the Indian Ocean, while preserving a consistent analysis of the remaining ocean basins in the Supplementary Material.



Revised Figure A1: Map of selected basins in this study. The study basins were created from the NOAA WOA23 mask file (which has a 1°x1° grid) ([website https://www.ncei.noaa.gov/data/oceans/woa/WOA23/MASKS/basinmask_01.msk](https://www.ncei.noaa.gov/data/oceans/woa/WOA23/MASKS/basinmask_01.msk)) then some basins were modified, in particular to divide them into North and South. 8 basins were defined and studied: North Indian (latitudinal limit at 0° S), South Indian (latitudinal limit at 40°S), North Atlantic (between 0° and 40° N), South Atlantic (latitudinal limit at 40°S), North Pacific, South Pacific, North Pacific with latitudes > 40°N, and Southern Ocean. North Atlantic is cropped at 40°N following the Caley et al. (2026) density-reconstruction method.

L542. Please comment on how this finding compares with your earlier findings.

We thank the reviewer for this suggestion. Following the revision process, this section has been substantially reorganized. Rather than discussing all ocean basins successively, the regional analysis now focuses exclusively on the North and South Indian Ocean, while the analyses of the Atlantic, Pacific and Southern Oceans have been moved to the Supplementary Material.

This new organization allows us to directly compare model performance between the two dynamically distinct sectors of the Indian Ocean using complementary skill metrics (RMSE, Bias, NSE and KGE), and to establish a clearer transition towards the subsequent discussion of the West–East density gradient. Consequently, the comparison with the global-scale evaluation has been rewritten accordingly throughout this section.

L546. Please provide more detail. The model's IQR are very small in the figures.

Fig. 10. I cannot see the IQR in the Southern Ocean panel, but the reason for this is never explained.

We thank the reviewer for highlighting this point. The very small IQR values in the Southern Ocean arise primarily from the limited number of available reconstruction points ($n = 7$), combined with the relatively homogeneous density values in this region. In contrast to other basins, Southern Ocean reconstructions are mostly located in open-ocean environments, with fewer coastal influences. This reduces the spatial variability captured by the proxy dataset, leading to a narrower distribution and thus a smaller IQR.

This figure was removed from the revised version of the manuscript and replaced with heatmaps of new statistical metrics, recommended by the second reviewer. These are available in the supplementary material (Figs. S5, S6 and S7).

L566-7. I cannot tell from the figure which are the ‘several regions’ you are referring to.

Right, the end of the sentence was unclear and has been deleted.

Section 4. Personally, I would remove this section (as it replicates Fig. 8 which I didn’t understand).

In agreement with the suggestion of the reviewer 1, we reduced this part by only treating in the main text the case of the Indian Ocean (North and South).

L708-710. This sentence reads awkwardly. I suggest removing it.

We agree with the reviewer and have removed this sentence for clarity.

Fig. 12. Can you please also add something to give readers an idea of how low-frequency variability in the IOD might compare to these changes in gradient.

We thank the reviewer for this valuable comment. We agree that it is important to contextualize the reconstructed West–East density gradient with respect to low-frequency

variability of the Indian Ocean Dipole (IOD). The key question is whether the simulations that best reproduce the mean West–East SST and density gradients in our study also simulate a realistic level of IOD variability.

We considered extending the discussion by directly comparing our results with the IOD variability analysis of Brierley et al. (2023). However, such a comparison is limited by the available model overlap. First, the simulations that best reproduce the reconstructed West–East gradient in our analysis are generally not included in Brierley et al. (2023). Second, the small number of overlapping simulations would not allow a robust assessment of the relationship between mean-state gradients and IOD variability, and, in our work, only 4 simulations have a correct mean state.

We have therefore expanded the discussion to clarify that the reconstructed LGM West–East density gradient reflects a shift in the mean state of the Indian Ocean, rather than a direct measure of interannual IOD variability. We also discuss perspectives about potential link between mean state changes and IOD variability in lines 871:

“This is of importance because mean state changes, as shown by Abram et al. (2020), likely enhanced interannual variability across the basin. Although previous study on IOD variability in PMIP simulations does not show a systematic relationship with changes in the mean zonal gradient across different climatic states in term of median value (Brierley et al., 2023), investigated if the simulations that perform the best to reproduce the gradient also show specific interannual variability would be interesting for future works.”

L784. This paper also highlight that HadCM3 was the most realistic, if I recall correctly

We thank the reviewer for pointing this out. We agree that previous studies have highlighted the strong performance of HadCM3 in simulating LGM climate patterns. In particular, DiNezio and Tierney (2013) showed that HadCM3 is the only model exhibiting statistically significant agreement with proxy-based rainfall reconstructions across the Indo-Pacific region.

In the revised manuscript, we expanded this discussion to place our results in the context of previous model evaluations. Our analyses confirm that the HadCM3 simulations perform particularly well in reproducing the reconstructed West–East density gradient in the Indian Ocean, consistent with their previously reported skill in simulating Indo-Pacific hydroclimate.

We added L.843:

“In particular, DiNezio and Tierney (2013) showed that HadCM3 is the only model exhibiting statistically significant agreement with proxy-based rainfall reconstructions in the Indo-Pacific region. In our analysis, some HadCM3 simulations also perform well, particularly in reproducing the west–east density gradient (Fig. 12).”

References added:

Caley, T., Rieger, N., Werner, M., Waelbroeck, C., Barathieu, H., Happé, T., and Roche, D. M.: Past Ocean surface density from planktonic foraminifera calcite $\delta^{18}\text{O}$, *Clim. Past*, 22, 247–263, <https://doi.org/10.5194/cp-22-247-2026>, 2026.

Caubel, A., Denvil, S., Foujols, M.-A., Marti, O., Dufresne, J.-L., Bopp, L., Cadule, P., Ethé, C., Idelkadi, A., Mancip, M., Masson, S., Mignot, J., Musat, I., Balkanski, Y., Bekki, S., Bony, S., Braconnot, P., Brockman, P., Codron, F., Cozic, A., Cugnet, D., Fairhead, L., Fichefet, T., Flavoni, S., Guez, L., Guilyardi, E., Hourdin, F., Ghattas, J., Kageyama, M., Khodri, M., Labetoulle, S., Lefebvre, M.-P., Levy, C., Li, L., Lott, F., Madec, G., Marchand, M., Meurdesoif, Y., Rio, C., Schulz, M., Swingedouw, D., Szopa, S., Viovy, N., and Vuichard, N.: IPSL-CM5A-LR model output prepared for CMIP5 piControl experiment, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.IPILpc, 2016.

Hajima, T., Abe, M., Arakawa, O., Suzuki, T., Komuro, Y., Ogura, T., Oguchi, K., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Ohgaito, R., Ito, A., Yamazaki, D., Ito, A., Takata, K., Watanabe, S., Kawamiya, M., and Tachiiri, K.: MIROC MIROC-ES2L model output prepared for CMIP6 CMIP piControl, Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.5710, 2019.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., Lipscomb, W. H., Long, M. C., Mahowald, N., Marsh, D. R., Neale, R. B., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W. D., Hack, J. J., Kiehl, J., and Marshall, S.: The Community Earth System Model: A Framework for Collaborative Research, *Bulletin of the American Meteorological Society*, 94, 1339–1360, doi: 10.1175/BAMS-D-12-00121.1, 2013.

JAMSTEC, AORI, and NIES: MIROC-ESM model output prepared for CMIP5 lgm, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.MIMElg, 2015.

JAMSTEC, AORI, and NIES: MIROC-ESM model output prepared for CMIP5 pi-Control, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.MIMEpc, 2015.

Jungclaus, J., Giorgetta, M., Reick, C., Legutke, S., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Fischer, N., Glushak, K., Gayler, V., Haak, H., Hollweg, H.-D., Kinne, S., Kornblueh, L., Matei, D., Mauritsen, T., Mikolajewicz, U., Müller, W., Notz, D., Pohlmann, T., Raddatz, T., Rast, S., Roeckner, E., Salzmann, M., Schmidt, H., Schnur, R., Segschneider, J., Six, K., Stockhause, M., Wegner, J., Widmann, H., Wieners, K.-H., Claussen, M., Marotzke, J., and Stevens, B.: cmip5 output2 MPI-M MPI-ESM-P lgm, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, <https://hdl.handle.net/21.14106/08f4f7f97af1441d977e14798e852166c5c35def>, 2014.

Jungclaus, J., Giorgetta, M., Reick, C., Legutke, S., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Fischer, N., Glushak, K., Gayler, V., Haak, H., Hollweg, H.-D., Kinne, S., Kornblueh, L., Matei, D.,

Mauritsen, T., Mikolajewicz, U., Müller, W., Notz, D., Pohlmann, T., Raddatz, T., Rast, S., Roeckner, E., Salzmänn, M., Schmidt, H., Schnur, R., Segschneider, J., Six, K., Stockhause, M., Wegner, J., Widmann, H., Wieners, K.-H., Claussen, M., Marotzke, J., and Stevens, B.: cmip5 output2 MPI-M MPI-ESM-P piControl, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, <https://hdl.handle.net/21.14106/ce74e48c66620583cc8c1fefb607de36a3d6ddb8>, 2014.

Kageyama, M., Denvil, S., Foujols, M.-A., Caubel, A., Marti, O., Dufresne, J.-L., Bopp, L., Cadule, P., Ethé, C., Idelkadi, A., Mancip, M., Masson, S., Mignot, J., Musat, I., Balkanski, Y., Bekki, S., Bony, S., Braconnot, P., Brockman, P., Codron, F., Cozic, A., Cugnet, D., Fairhead, L., Fichefet, T., Flavoni, S., Guez, L., Guilyardi, E., Hourdin, F., Ghattas, J., Khodri, M., Labetoulle, S., Lefebvre, M.-P., Levy, C., Li, L., Lott, F., Madec, G., Marchand, M., Meurdesoif, Y., Rio, C., Schulz, M., Swingedouw, D., Szopa, S., Viovy, N., and Vuichard, N.: IPSL-CM5A-LR model output prepared for CMIP5 lgm experiment, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.IPILlg, 2016.

NASA/GISS: NASA-GISS: GISS-E2-R model output prepared for CMIP5 last glacial maximum, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.GIGRIg, 2014.

NASA/GISS: NASA-GISS: GISS-E2-R model output prepared for CMIP5 pre-industrial control, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.GIGRpc, 2014.

Ohgaito, R., Abe-Ouchi, A., Abe, M., Arakawa, O., Ogochi, K., Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Ito, A., Yamazaki, D., Ito, A., Takata, K., Watanabe, S., Kawamiya, M., and Tachiiri, K.: MIROC MIROC-ES2L model output prepared for CMIP6 PMIP lgm, Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.5644, 2019.

Sénési, S., Richon, J., Franchistéguy, L., Tyteca, S., Moine, M.-P., Voldoire, A., Sanchez-Gomez, E., Salas y Mélia, D., Decharme, B., Cassou, C., Valcke, S., Beau, I., Alias, A., Chevallier, M., Déqué, M., Deshayes, J., Douville, H., Madec, G., Maisonnave, E., Planton, S., Saint-Martin, D., Szopa, S., Alkama, R., Belamari, S., Braun, A., Coquart, L., and Chauvin, F.: CNRM-CM5 model output prepared for CMIP5 lgm, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.CEC5lg, 2014.

Sénési, S., Richon, J., Franchistéguy, L., Tyteca, S., Moine, M.-P., Voldoire, A., Sanchez-Gomez, E., Salas y Mélia, D., Decharme, B., Cassou, C., Valcke, S., Beau, I., Alias, A., Chevallier, M., Déqué, M., Deshayes, J., Douville, H., Madec, G., Maisonnave, E., Planton, S., Saint-Martin, D., Szopa, S., Alkama, R., Belamari, S., Braun, A., Coquart, L., and Chauvin, F.: CNRM-CM5 model output prepared for CMIP5 piControl, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.CEC5pc, 2014.

Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Tanaka, T., Shindo, E., Tsujino, H., Deushi, M., Mizuta, R., Yabu, S., Obata, A., Nakano, H., Koshiro, T., Ose,

T., and Kitoh, A.: MRI-CGCM3 model output prepared for CMIP5 lgm, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.MRMClg, 2015.

Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Tanaka, T., Shindo, E., Tsujino, H., Deushi, M., Mizuta, R., Yabu, S., Obata, A., Nakano, H., Koshiro, T., Ose, T., and Kitoh, A.: MRI-CGCM3 model output prepared for CMIP5 piControl, served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.MRMCpc, 2015.

Reviewer n°2

The manuscript “Using ocean surface paleo-density to evaluate PMIP3 1 and PMIP4 Last Glacial Maximum climate simulations” provides a novel data-model comparison for the LGM using a new surface density reconstruction. Surface density is the integration of both sea surface temperature and sea surface salinity, which allows for a more robust evaluation of model performance that indirectly includes the model representation of changes in the hydrological cycle (via changes in coastal salinity). The authors include an analysis of SST versus data for additional context, where the MARGO database and Tierney et al. (2020) data sets were used as a benchmark. Model performance of SSS was not directly benchmarked. While I feel the analysis provides valuable insight into the PMIP simulations, there is also a reliance on R^2 correlation and regression slopes, which are likely imperfect metrics. These may still be acceptable choices as finding the appropriate statistical methods to compare data and model under paleo settings itself is a non-trivial task. As it is used herein, correlation is a rough estimate that the spatial patterns of data LGM minus LH match those of the model simulations. The regression slopes between data anomalies and model anomalies loosely test if the magnitude of the model anomalies match the data, within the context of the spatial patterns, where the authors used a 0.8 to 1.2 (+/- 20%) bound. As an example, one could simply add +5C to the data globally, and the correlation would have an R^2 of 1 and a slope of 1. The authors may choose to address this near line 478. I found the distribution-based KS test and IRQ-overlap evaluations to be more satisfying, so I’m unsure why the authors didn’t apply those methods to the raw values (LGM: data vs model) in addition to the anomalies (LGM-LH model vs data). More rigorous metrics, such as RMSE, could have been used to evaluate model skill, but these metrics may be unkind to the models in raw terms.

Overall the text is well written and highlights valuable findings using a novel approach, but I am unsure on the overreliance on regression and correlation as the primary metrics of model

skill without additional justification or consideration of other metrics, such as RMSE, Nash–Sutcliffe efficiency (NSE), or Kling–Gupta efficiency (KGE).

We thank Reviewer 2 for their careful and constructive assessment of our manuscript. We appreciate the recognition of the novelty and relevance of our surface density approach. We also acknowledge the concerns regarding the reliance on correlation and regression metrics, and we have revised the manuscript accordingly to include additional, more robust statistical diagnostics. Below, we present a point-by-point response to the individual comments raised. The reviewer's comments are in black, and our responses are **in purple**.

“There is a reliance on R^2 correlation and regression slopes, which are likely imperfect metrics... one could simply add $+5^\circ\text{C}$ to the data globally and still obtain $R^2 = 1$ and slope = 1.”

We thank the reviewer for this important comment. We agree that correlation (R^2) and regression slopes provide only partial information on model simulation performance and may not fully capture biases in absolute values. As correctly pointed out, correlation primarily reflects the similarity in spatial patterns, while regression slopes provide a first-order estimate of amplitude agreement, but both metrics can be insensitive to systematic offsets.

To address this limitation, we have complemented our analysis with additional performance metrics, namely Root Mean Square Error (RMSE), Nash–Sutcliffe Efficiency (NSE), and Kling–Gupta Efficiency (KGE). These metrics provide a more comprehensive evaluation of model simulation skill by explicitly accounting for bias, variability, and overall agreement with observations.

We have added a methodological description of these metrics in Section 2.3 (Statistical analysis), as follows L.276:

“To further compare data and model simulations and not overreliance on regression and correlation, we also rely on other metrics. Root mean square error (RMSE) quantifies the average magnitude of errors between model and observations (lower values indicate better agreement), while bias measures the mean signed difference and thus highlights systematic over- or under-estimation (values close to zero are optimal). The Nash–Sutcliffe efficiency (NSE) evaluates the ability of model simulations to reproduce the observed variability. Values close to 1 indicate high skill, whereas negative values indicate poor model performance. The Kling–Gupta efficiency (KGE) combines correlation, variability, and bias into a single metric, providing an integrated measure of model performance, with values close to 1 indicating better agreement.”

In addition, we have substantially revised the global evaluation (Sect. 3.3). Rather than relying primarily on regression statistics, model performance is now assessed using a combination of Taylor diagrams, complementary performance metrics (RMSE, Bias, NSE and KGE), and regression plots. While Taylor diagrams summarize spatial correlation, normalized standard deviation and centered root-mean-square error (RMSE), the new metrics explicitly quantify

systematic bias, error magnitude and overall model skill. The regression plots are now used only as a visual complement illustrating the point-by-point relationship between reconstructed and simulated surface densities.

These complementary metrics are presented in the revised Figure 9, 10 and 11, and constitute the main framework used to discuss model performance throughout Sect. 3.3 and sect.4.1. We therefore believe that the revised multi-metric approach provides a much more robust assessment of model skill than relying solely on R^2 and regression slopes as suggested by reviewer 2. All the associated discussion in parts 3.3 and 4.1 have been completely revised to include these new metrics and interpretation.

Part 3.3:

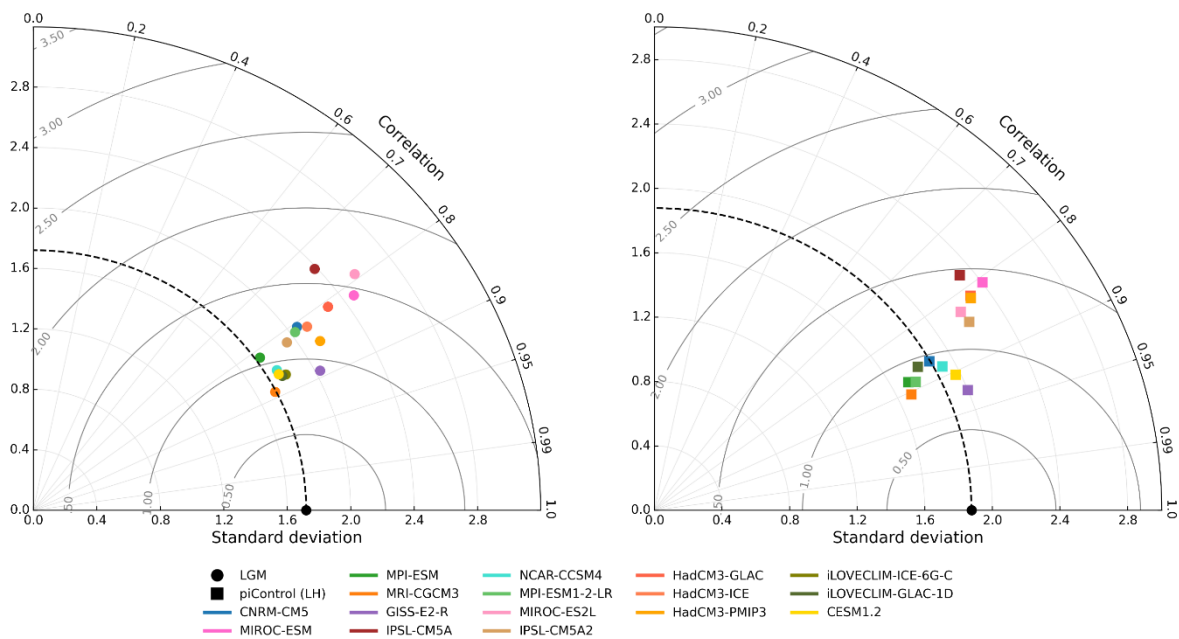


Figure 8: Taylor diagrams comparing model simulations performance for LGM (left) and piControl (LH, right) ocean surface density reconstructions. All grid cells are equally weighted in these diagrams. Each point represents a climate model simulation. Circles correspond to the Last Glacial Maximum (LGM), while squares correspond to the pre-industrial control simulation (piControl, LH). The radial distance indicates the model simulation standard deviation, while the angular coordinate represents the pattern correlation between model outputs and reconstructions. The black dots define the reference observational standard deviation which represents the spatial variability of the observational reconstructions, computed as the standard deviation of all observations within each period. Contours show centered root-mean-square error (CRMSE) relative to observations.

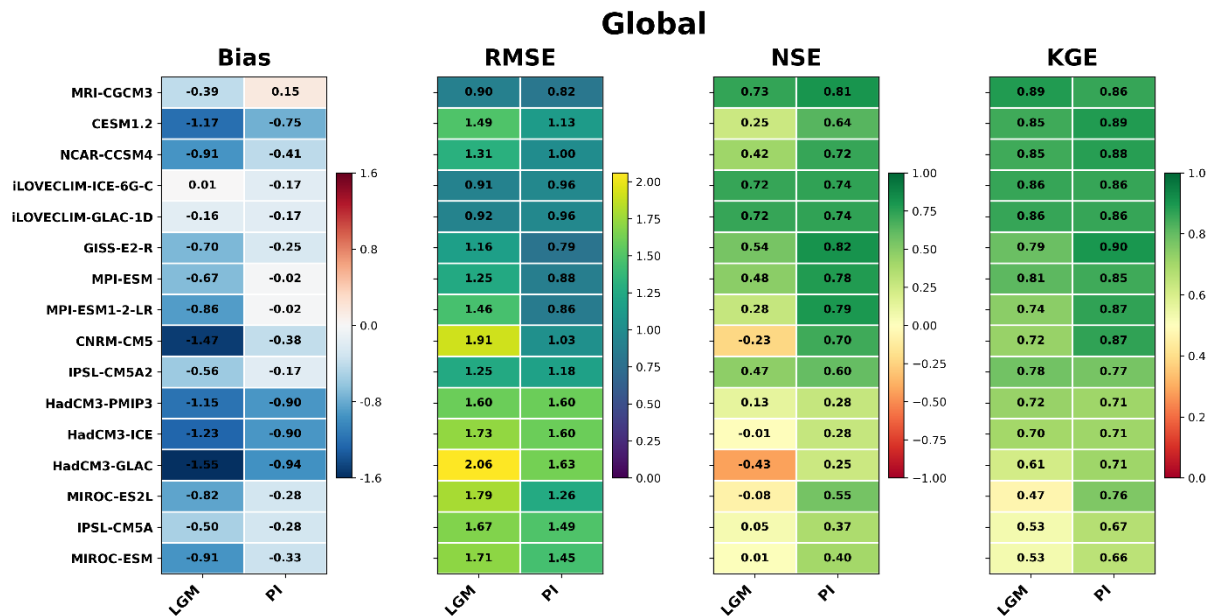


Figure 9: Global model simulations performance for sea surface density during the LGM and PI periods. Heatmaps show four complementary metrics calculated from model–data comparisons at reconstruction grid cells common to all simulations: bias, root mean square error (RMSE), Nash–Sutcliffe efficiency (NSE), and Kling–Gupta efficiency (KGE). 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.

Part 4 and 4.1:

“ 4. Evaluate model simulations at regional scale: focus on the Indian Ocean

Despite global-scale agreement between reconstructions and simulations, the mismatches observed with statistical metrics reveals that key patterns are not captured. This motivates a basin-by-basin analysis to better understand the regional origins of these discrepancies.

We created 8 ocean basins masks (see Fig. A1). To evaluate model performance within coherent oceanographic regions while accounting for the latitudinal structure identified in the global analysis, oceans were subdivided into northern and southern sectors (Appendix A). In the following, we focus the main discussion on the Indian Ocean. This choice is motivated by the central role of the Indian Ocean in tropical–extratropical interactions, it is a place of major hydroclimate processes such as monsoon and Walker circulation dynamics and previous studies already conducted some data model comparison in this region, enlightening its complexity.

Analyses of the other ocean basins are provided in the Supplementary Material (Figs. S5, S6 and S7) using three statistical metrics (R, RMSE and Bias). Detailed analyses of these ocean basins will be for future studies but, briefly, performance of the data model comparison is variable depending of the ocean basin and simulations considered. The South Atlantic and North Pacific (north of 40°N) ocean basins indicate the highest bias and RMSE and lowest correlation for weighted averages simulations.

Regional model simulations performance for the Indian Ocean basin are evaluated using complementary skill metrics (Bias, RMSE, NSE and KGE), providing a more direct assessment of model–data agreement.”

Part 4.1:

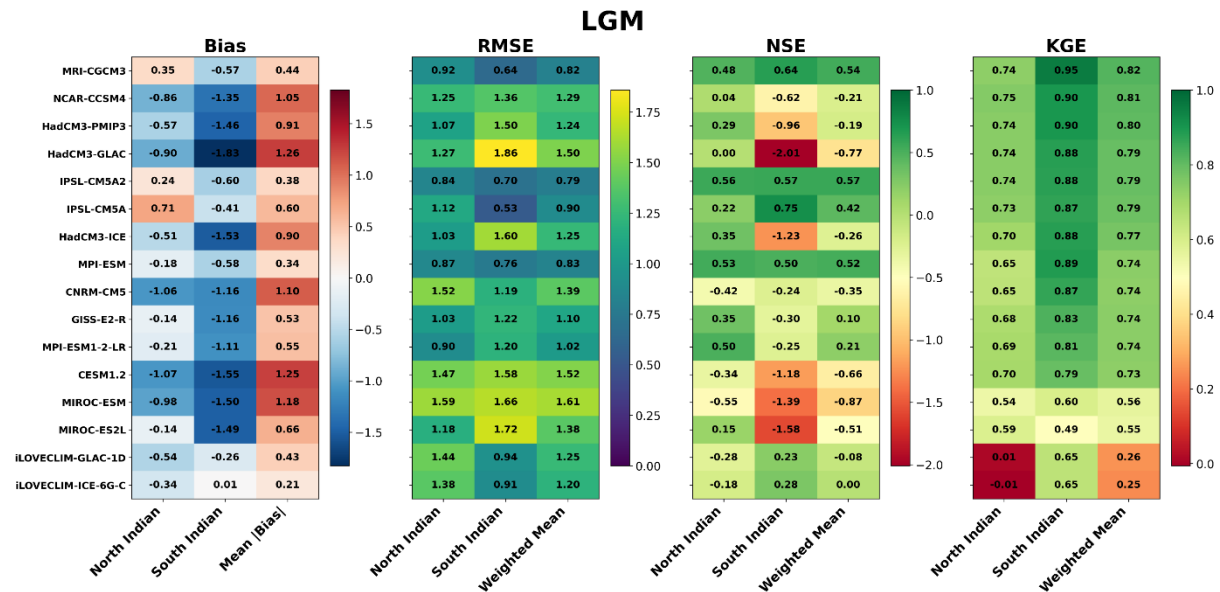


Figure 10: Simulations performance for seawater density in the North Indian and South Indian oceans during the Last Glacial Maximum (LGM). Heatmaps show four complementary performance metrics calculated from model–data comparisons at proxy reconstruction sites. Models are ranked according to their weighted mean Kling–Gupta Efficiency (KGE), averaged across both basins. The metrics shown are: mean absolute bias ($|Bias|$), 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 $|Bias|$ and RMSE indicate better agreement with the reconstructions, whereas higher values of NSE and KGE indicate better model simulations performance, with an optimum value of 1.

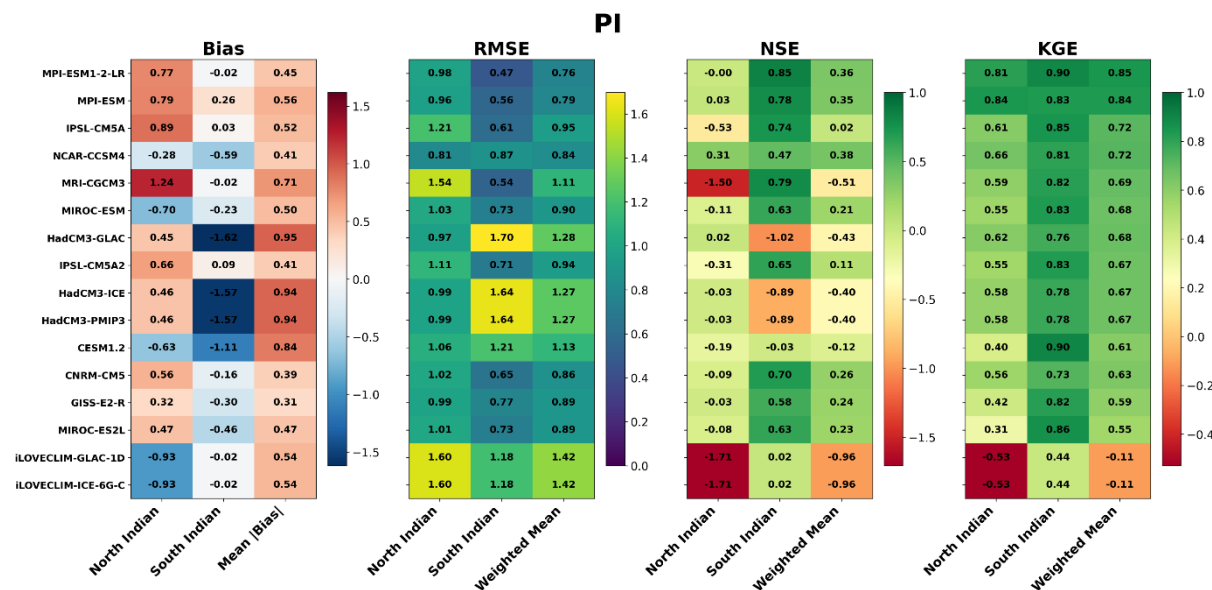


Figure 11: Simulations performance for seawater density in the North Indian and South Indian oceans during the pre-industrial period (PI). Heatmaps show four complementary performance metrics calculated from model–data comparisons at proxy reconstruction sites. Models are ranked according to their weighted 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, whereas higher values of NSE and KGE indicate better model simulations performance, with an optimum value of 1.

“More rigorous metrics, such as RMSE, NSE, or KGE, could have been used...”

We thank the reviewer for this suggestion. Following this recommendation, we have incorporated additional performance metrics, including RMSE, bias, NSE, and KGE, to complement the correlation-based diagnostics. Thanks to the reviewer n°1 we also added Taylor diagrams for the global scale.

We note that while these metrics provide a stricter evaluation of model simulation performance in absolute terms, they can be sensitive in paleo-contexts where proxy uncertainties are large and spatial sampling is sparse.

For this reason, we adopt a multi-metric approach combining correlation-based, distribution-based, and error-based diagnostics to provide a balanced and robust assessment of model simulations skill.

Minor comments:

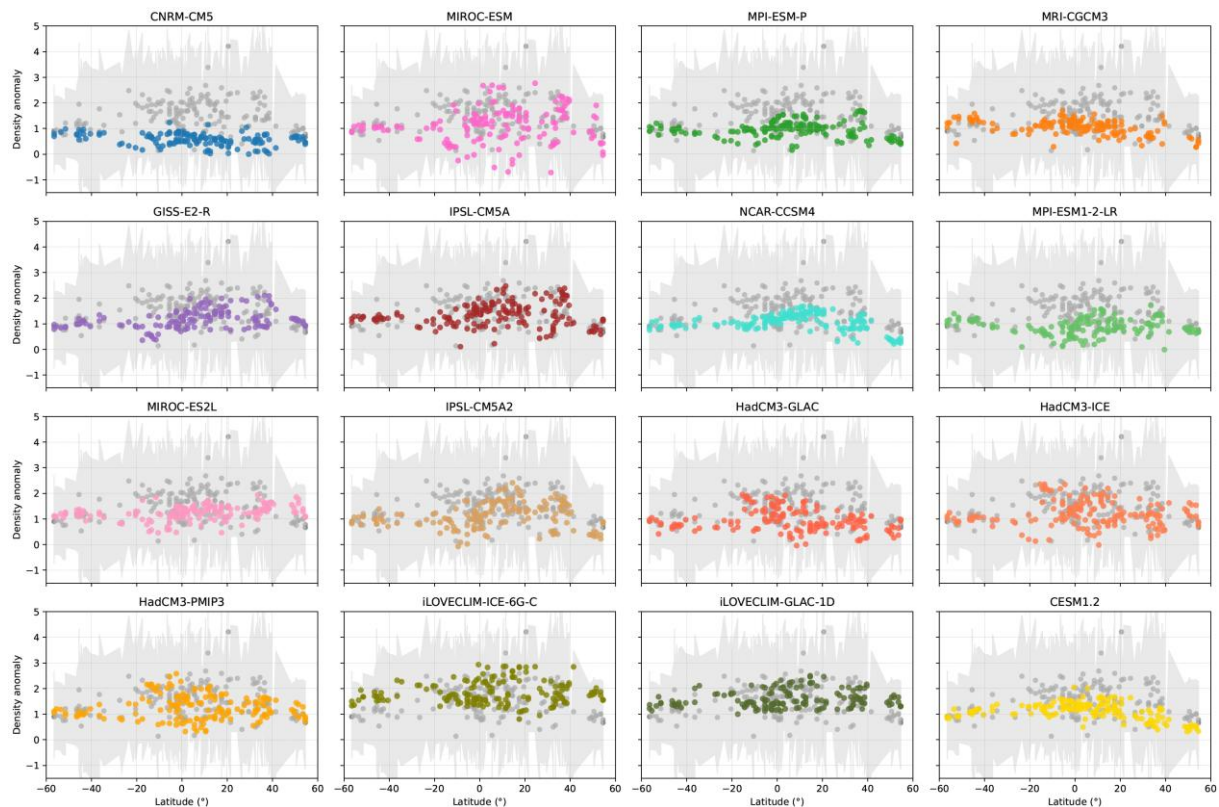
Line ~236: note the global KS threshold for the sample size.

We add this sentence L.253:

“At the global scale, the KS threshold is 0.13, consistent with the sample size of our dataset.”

Figure 3, MPI-ESM-P label isn’t in the plot.

We thank the reviewer for this comment. The model simulation label has been corrected to “MPI-ESM-P” in Figure 3, as well as in Figures B1–B4, to ensure consistency across all figures.



Revised Figure 3: Density anomaly (kg/m³) as a function of latitude for each model simulation (colored dots) compared with the observational data (grey dots) and the 95% confidence interval (grey shading). Model outliers, identified using the interquartile range (IQR) method (values outside 1.5×IQR), were excluded to reduce the influence of extreme values. This filtering highlights the main structure and latitudinal patterns of the modelled density anomalies while retaining all available latitude points.

Figure 7, model median is in the legend, but I don't think it is shown in the plot.

In Figure 7, the model simulation median is indeed represented by the circular markers for each simulation (blue).

Line 445: Does figure 7 show all three criteria?

We thank the reviewer for this question, which allowed us to clarify the respective roles of Figures 6 and 7.

Figure 7 only explicitly shows the median and interquartile range (IQR) of the reconstructed and simulated distributions. The Kolmogorov–Smirnov (KS) statistic and associated p-value are not displayed graphically in this figure but are evaluated numerically and summarized through the coloured flags shown in Figure 6.

During the revision process, this analysis was further improved by propagating reconstruction uncertainties through a Monte Carlo framework according to reviewer 1 suggestion. Consequently, Figure 7 has also been updated and now displays both the original

reconstructed IQR and the Monte Carlo-propagated reconstructed IQR, together with their corresponding medians. However, as in the previous version, only the IQR criterion is represented in Figure 7, whereas the Monte Carlo-derived KS statistic and p-value remain summarized through the coloured flags in Figure 6.

We revised the corresponding paragraph in Sect. 3.2 to make this distinction explicit.

Figure 12: it is not clear to me why the gray uncertainty bands don't align to the data marker (star + 1-sigma).

We thank the reviewer for this comment. The grey uncertainty bands represent the standard deviation of the model simulations ($\pm 1\sigma$ and $\pm 2\sigma$), while the star markers correspond to reconstruction-based anomalies. These two elements are therefore not directly centred on one another, as they do not represent the same underlying quantity. We have clarified this distinction in the figure caption to avoid confusion.

Figure E1: In the caption, note LH = Late Holocene

Done.

“Figure E1: Boxplots showing West–East anomalies (Indian_East – Indian_West) in Sea Surface Temperature (°C), (from the MARGO and Tierney datasets, separately), Sea Surface Salinity (g/kg), and sea surface density (kg/m³) during the pre-industrial period. The boxplots represent the distribution of anomalies from climate model simulations only. Each colored dot corresponds to the anomaly from an individual model simulation. Red stars indicate observed values from the EN4 dataset (Good et al., 2013), averaged over the period 1900–1999 to provide a historical reference. Orange stars indicate reconstructions from databases (PI from WOA18 for MARGO (MARGO project, 2009), LH for Tierney et al. (2020) and LH (**Late Holocene**) for density (Caley et al.2026)).”

~Line 730: Why do the GCM SST anomalies vary so much between Tierney (Fig 12b) and MARGO (Fig 12c). Is it from the location of the sites or the differing counts?

We thank the reviewer for this question. The differences in SST anomalies between Tierney et al. (Fig. 12b) and the MARGO compilation (Fig. 12c) arise from several factors. First, the two datasets are based on different sets of core locations, leading to distinct spatial sampling of the ocean. Second, the number of available records differs between the two compilations, which affects the robustness of regional averages. Finally, the reconstructions rely on different proxy calibrations and methodological frameworks, which can also lead to systematic differences in estimated SST anomalies. Together, these factors explain the observed discrepancies.