Response to reviewer 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 will 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 will reduce 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:

[…] New and published δ^{18}Oc 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 $\delta^{18}O_c$ 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).

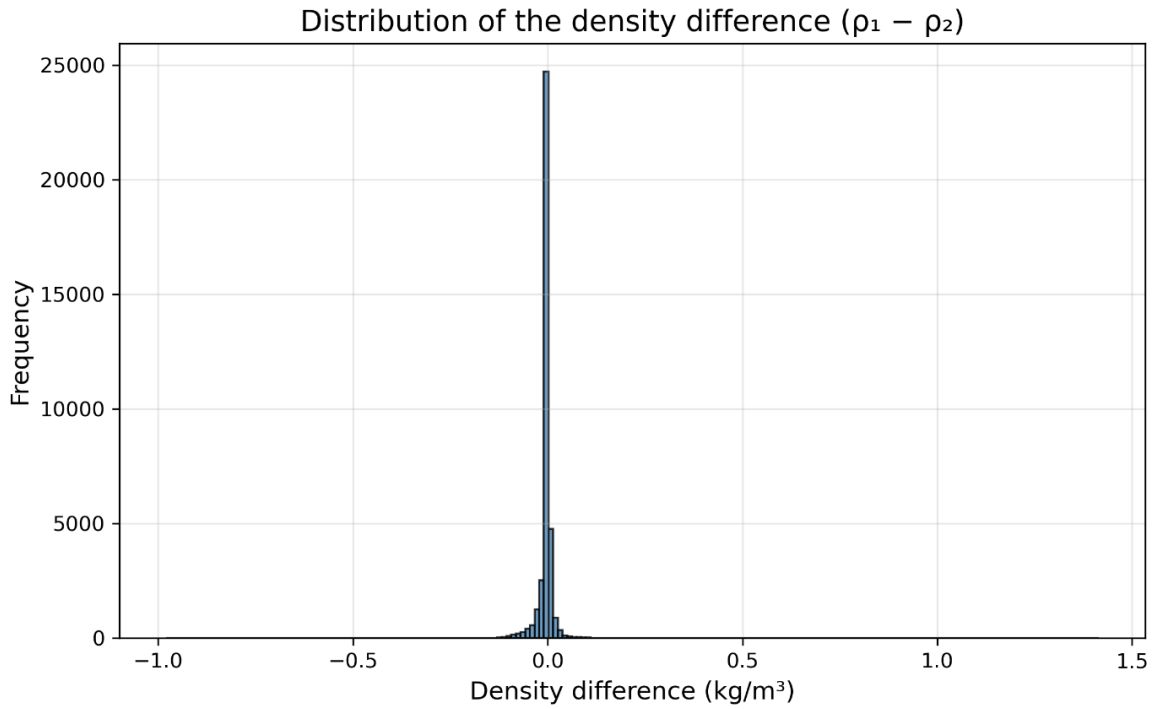


Figure A: Histogram example of the surface density difference between the two approaches for GISS-E2-R_lgm simulation (ρ_1: density computed with the original resolution and then regridded at 1°x1°, ρ_2: density recomputed from regridded SST/SSS at 1°x1°). The distribution is strongly peaked around zero, 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 ESFG?

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. Also 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.

We have added a clarification in the revised manuscript after L241:

"Here, *n* corresponds to the number of grid cells with available observations in each basin. This sample size is used to define the basin-specific KS critical threshold, which varies across basins because the KS statistic depends on sample size."

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:

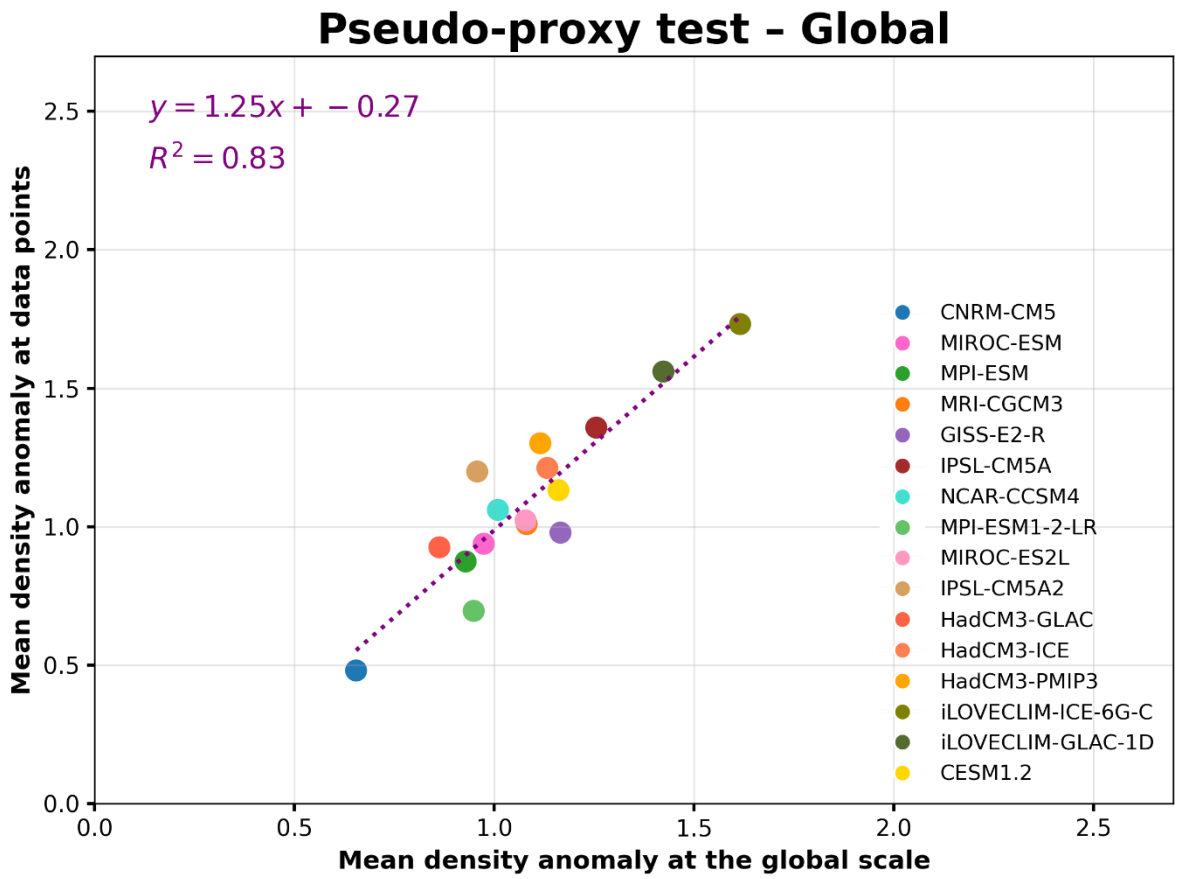
"Here, the pseudo-proxy approach refers to a spatial representativeness test: model simulations values are sampled at the exact locations of the proxy reconstructions and compared with the basin-wide model simulations mean. This allows us to assess whether the uneven proxy network provides an unbiased estimate of the large-scale basin signal."

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

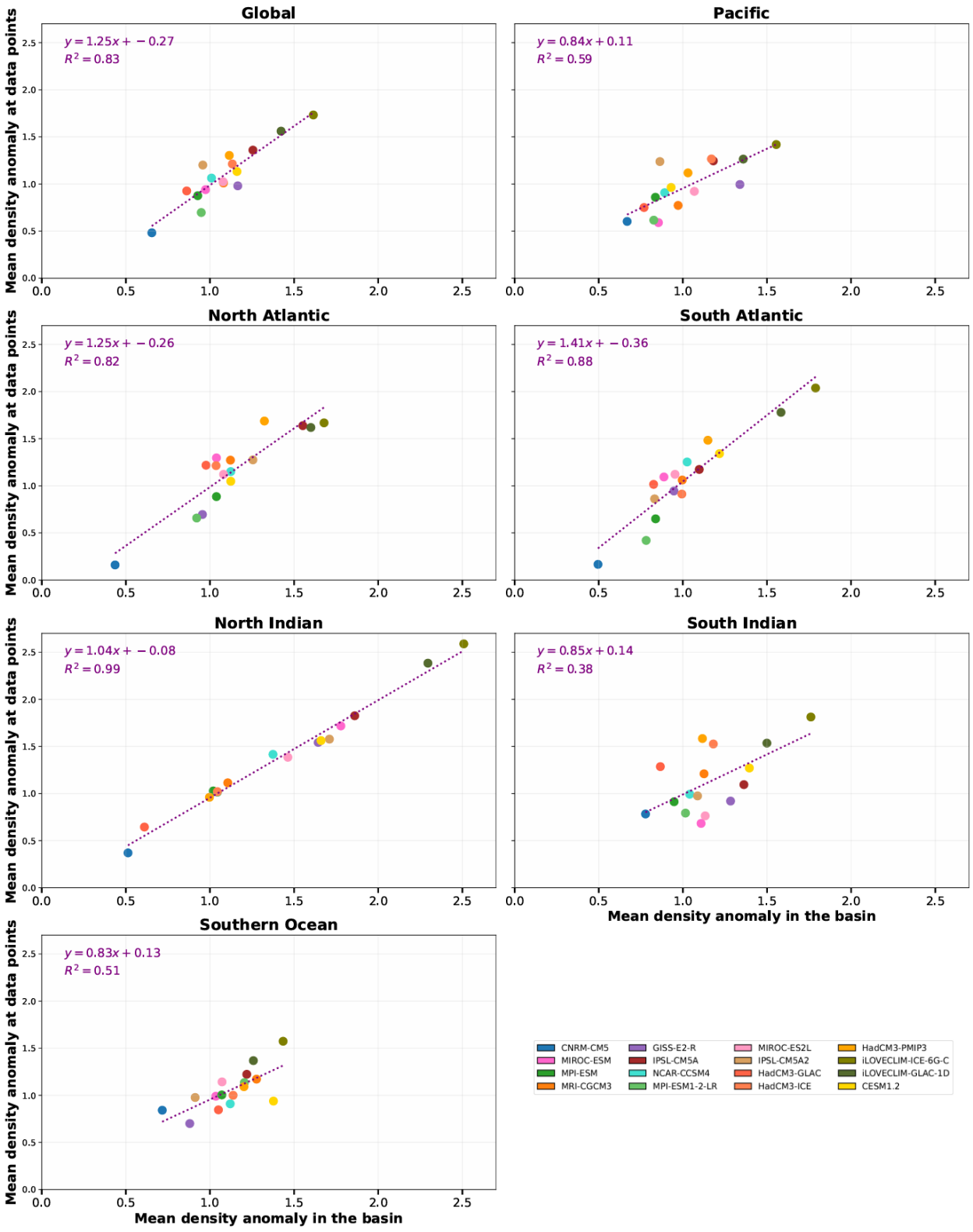
Actually, there was a mistake in how this x-axis was labelled; it is indeed the global axis, and this will be corrected. Thank you to have bringing this to our attention.

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 rest of the manuscript does not rely on a PMIP3/PMIP4 separation, 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/m3) 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^2 represent the fit across all model simulations. Results for individual basins are provided in Supplementary Fig.S2.



Revised Figure S2: Pseudo-proxy test for 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^2 represent the fit across all model simulations. 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 density anomaly driven only by regional hydrographic changes in SST and SSS (without global ice-sheets effect on global ocean salinity) (See also Caley et al., 2026 for detail explanations). 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/m³ density correction to the reconstructions, following Caley et al. (2026)."

L308. Please put the values in Table 1.

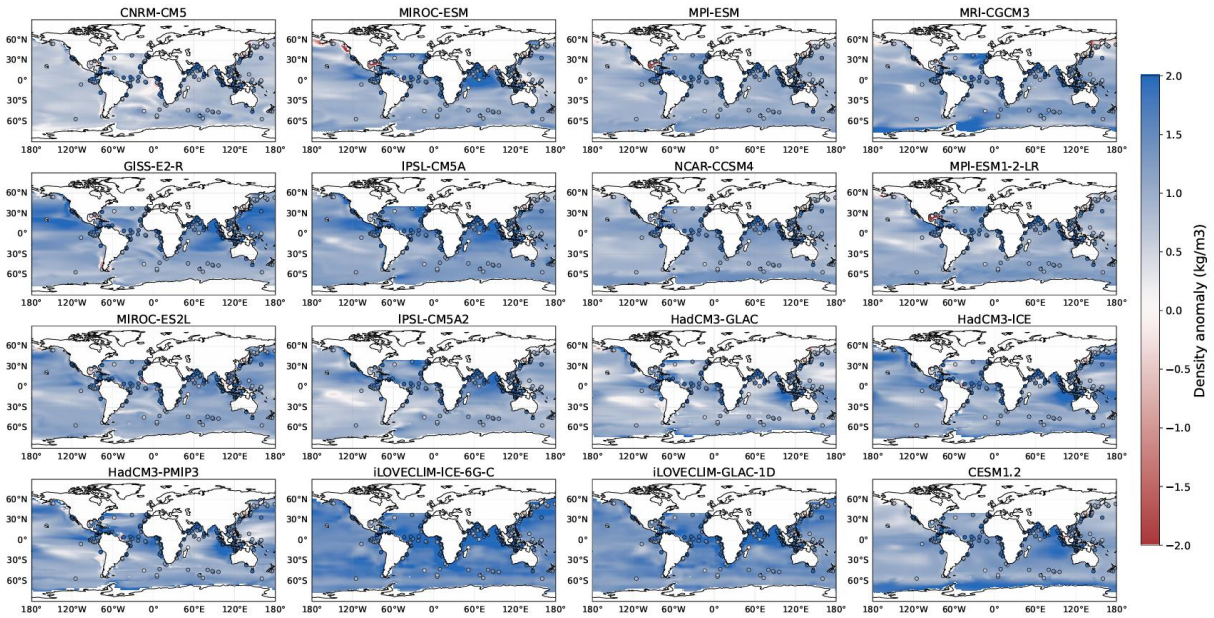
We thank the reviewer for this suggestion. We have computed the mean LGM–PI surface density anomaly for each model simulation, using a latitude-weighted spatial average, and we now include these values in Table 1, as individual model simulations estimates, and as an ensemble mean over all simulations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

Model simulation	PMIP/CMIP - Spin-up phase	Mean LGM-PI surf. density anomaly ($kg\,m^{-3}$)	References
CNRM-CM5	PMIP3-CMIP5 r1i1p1	0.672	Voldoire et al. (2013)
MIROC-ESM	PMIP3-CMIP5 r1i1p1	1.002	Sueyoshi et al. (2013)
MPI-ESM	PMIP3-CMIP5 r1i1p1	0.924	Adloff et al. (2018)
MRI-CGCM3	PMIP3-CMIP5 r1i1p1	1.055	Yukimoto et al. (2012)
GISS-E2-R	PMIP3-CMIP5 r1i1p150	1.269	Schmidt et al. (2014) and Ullman et al. (2014)
IPSL-CM5A	PMIP3-CMIP5 r1i1p1	1.357	Dufresne et al. (2013)
NCAR-CCSM4	PMIP3-CMIP5 r1i1p1	0.998	Gent et al. (2011) and Brady et al. (2013)
HadCM3B-M2.1aD (PMIP3 ice sheets)	PMIP4-CMIP6	1.172	Izumi et al. (2023)
HadCM3B-M2.1aD (GLAC-1D)	PMIP4-CMIP6	0.929	Izumi et al. (2023)
HadCM3B-M2.1aD (ICE-6GC)	PMIP4-CMIP6	1.211	Izumi et al. (2023)
MPI-ESM1-2-LR	PMIP4-CMIP6 r1i1p1f1	0.919	Mauritsen et al. (2019)
MIROC-ES2L	PMIP4-CMIP6 r1i1p1f2	1.115	Hajima et al. (2020)
IPSL-CM5A2	PMIP4-CMIP6	1.058	Sepulchre et al. (2020)
iLOVECLIM1-1-1-GLAC-1D	PMIP4-CMIP6	1.450	Lhardy et al. (2021) and Bouttes et al. (2023)
iLOVECLIM1-1-1-ICE-6G-C	PMIP4-CMIP6	1.641	Lhardy et al. (2021) and Bouttes et al. (2023)
CESM1-2	PMIP4-CMIP6	1.215	Tierney et al. (2020)

Revised Table 1: Model simulations available for this study, with LGM and piControl simulations. Model simulations from PMIP3 (blue) and PMIP4 (pink) and mean LGM-PI surface density anomaly for each simulation (kg/m^3)

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.

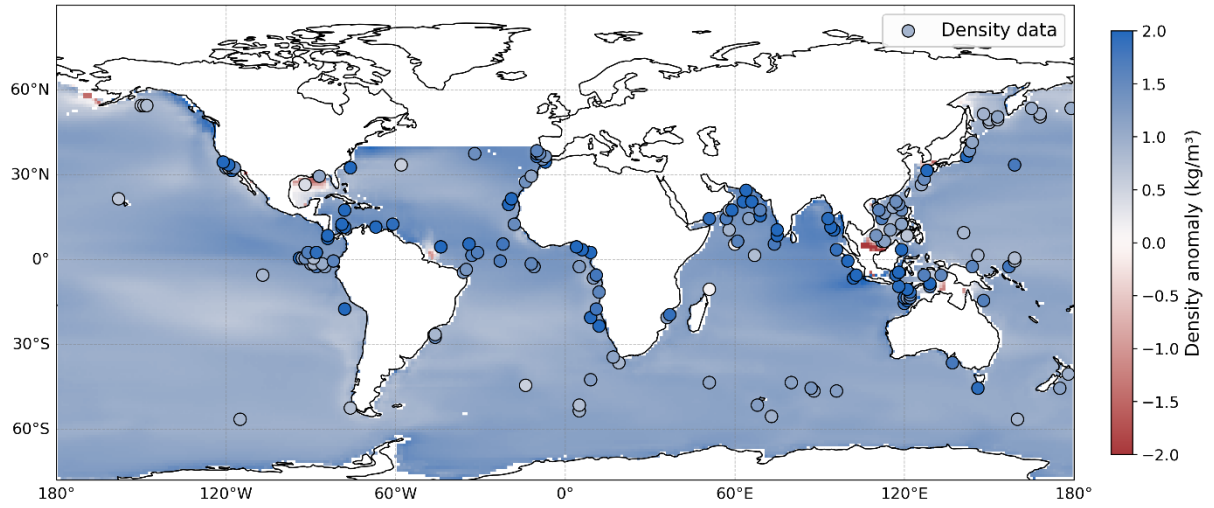
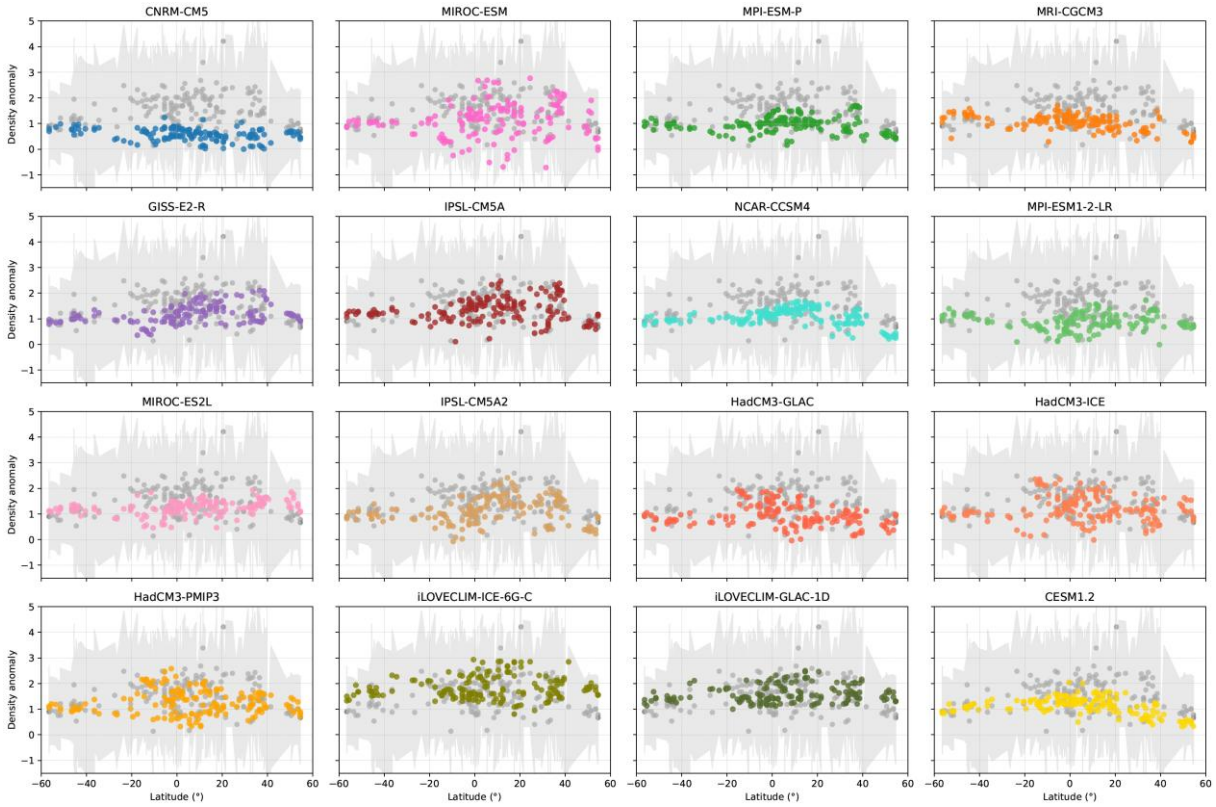


Figure S1 (new): Absolute surface density (kg/m^3) anomaly map (LGM - piControl). The dots represent the surface density anomaly database reconstructions (Caley et al., 2026) and the background map is the anomaly mean for all model simulations in the study.

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×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?

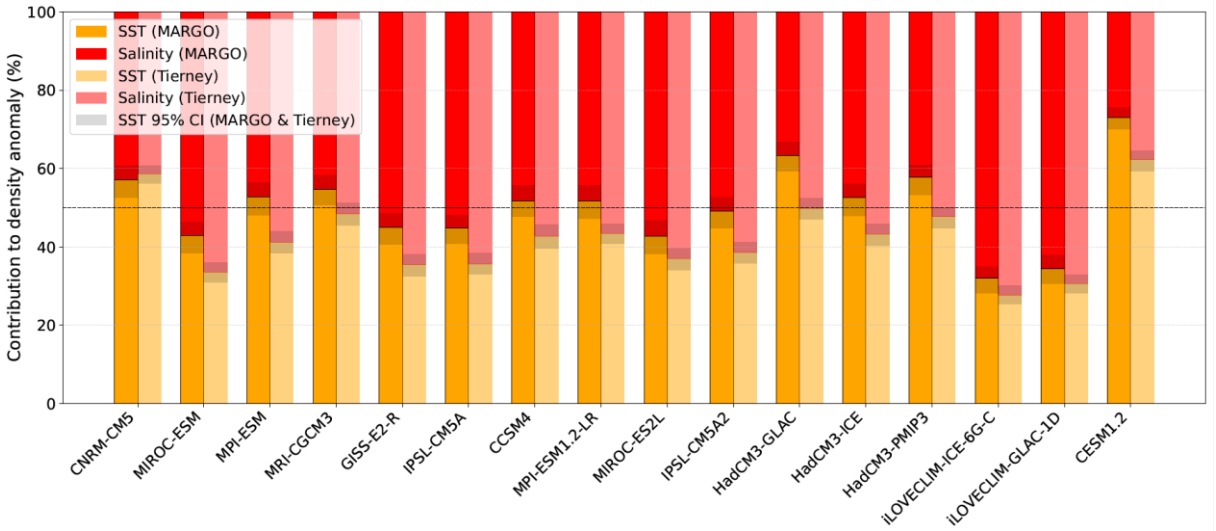
We thank the reviewer for these important comments.

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 ~10^{-4} kg/m³), 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 after L.233:

“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.”

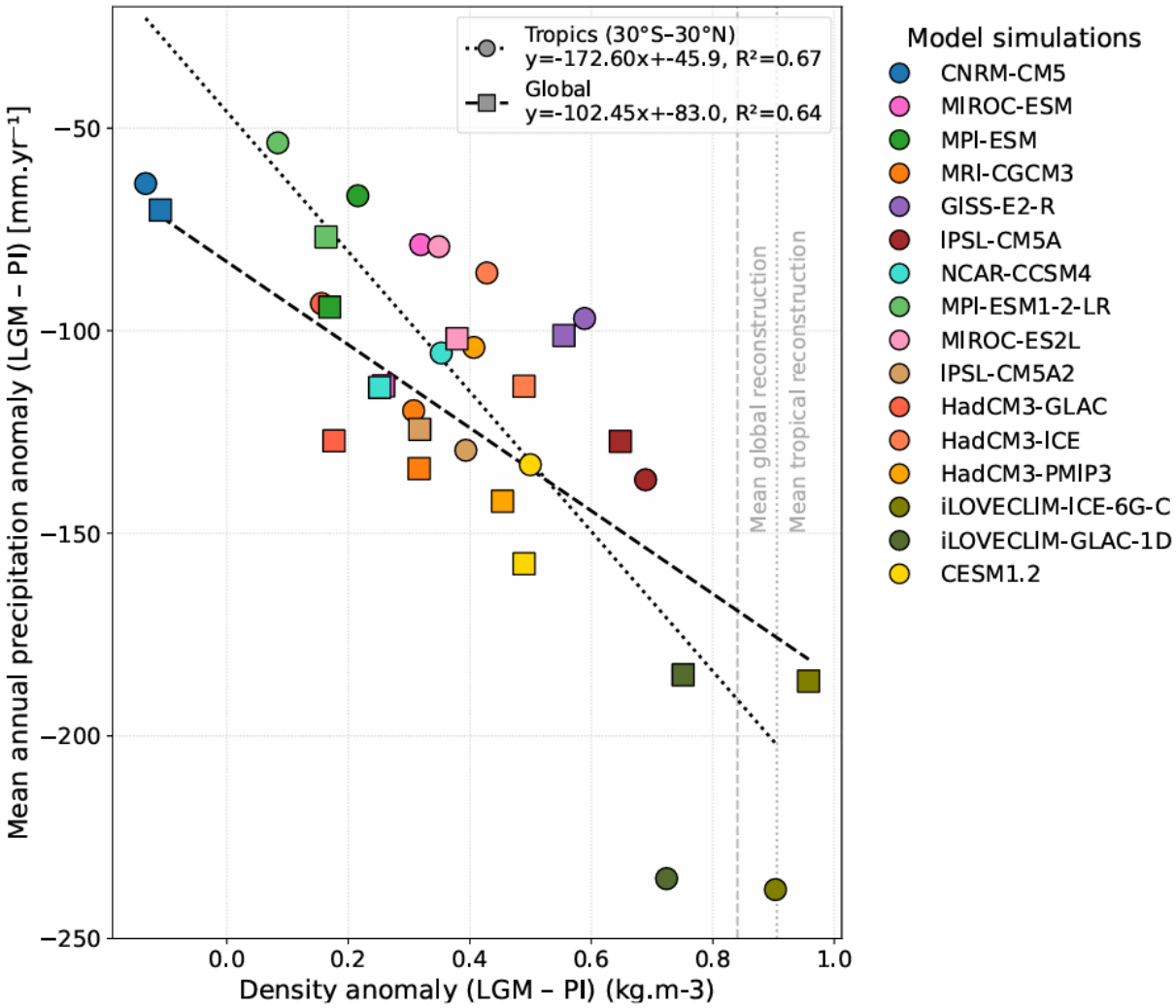
Fig. 5. Previously you stated the global reconstruction mean of the density anomaly was 1.5 (L310), but here it is shown as 0.8. Please clarify!

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/m^3) 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 R^2. 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 important comment. The IQR values reported in the manuscript are currently based on the spatial distribution of the mean LGM–PI density anomalies across reconstruction sites and do not explicitly account for calibration uncertainty through, for example, a Monte Carlo propagation approach.

We agree that including calibration uncertainty could affect the spread of the reconstructed distribution and, consequently, influence some of the criteria used to assess model–data agreement. We are currently performing additional tests to evaluate the impact of

incorporating calibration uncertainty (e.g. using the residual variance from the proxy calibration) on the IQR and associated diagnostics.

The manuscript will be revised accordingly to reflect these results and to clarify how this source of uncertainty affects our conclusions.

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 a useful assessment of how well the model simulation reproduces the spatial variability 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 variability (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 plots (below) and we also included additional statistical parameters as suggested by Reviewer 2 (see response to Reviewer 2).

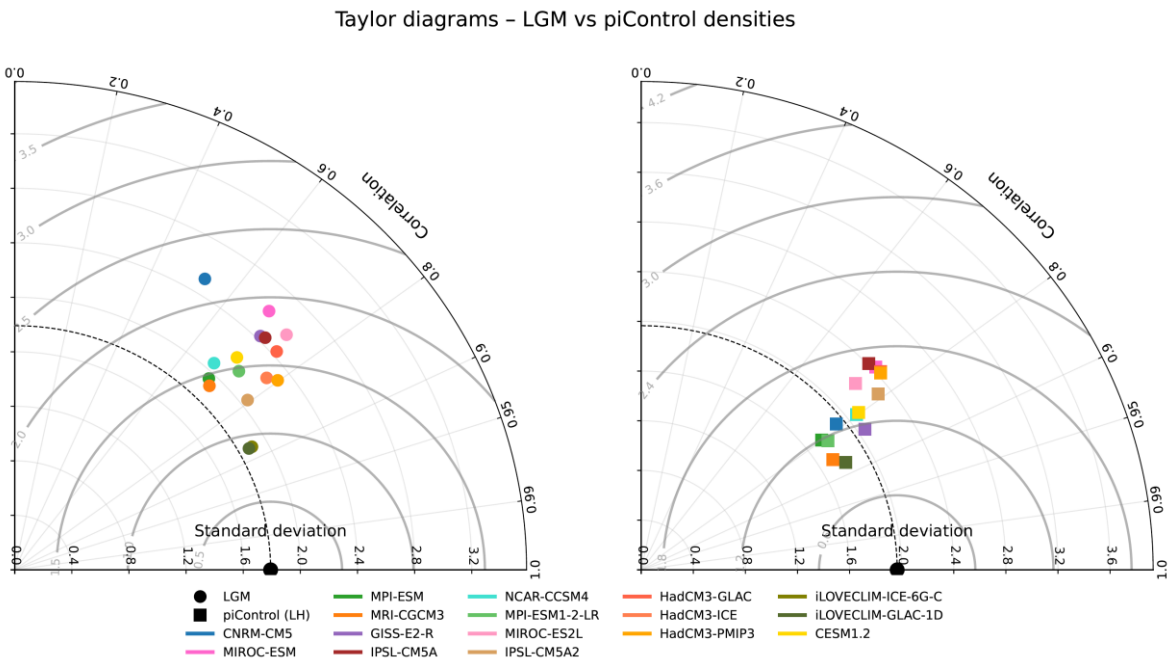


Figure B: Taylor diagrams comparing model simulations performance for LGM (left) and piControl (LH, right) ocean surface density reconstructions. All grid cells are equally weighted in these diagrams. Each point represents a climate model simulation. Circles correspond to the Last Glacial Maximum (LGM), 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.

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 helps us clarify the idea behind our regional decomposition.

In the previous section, we show that tropical and extratropical responses differ in terms of density anomalies and associated hydrological changes. However, extending this separation systematically within each ocean basin would substantially increase the complexity of the analysis and dilute the interpretability of basin-scale circulation controls.

In this revised version, we further enhance the analysis by focusing the main text on the Indian Ocean. This region is crucial for tropical–extratropical interactions and has a high density of reconstructions. 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 will be 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.

We also choose to focus the main discussion on the Indian ocean to make an easier connection with the last part of our discussion dealing with the West-East gradient.

The analysis of other basins (Atlantic, Pacific and Southern Ocean) will be provided in 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. We will also include a supplementary figure explicitly comparing tropical versus extratropical contributions across all basins.

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

We thank the reviewer for this suggestion. We now explicitly compare these results with our earlier global-scale findings. At the global scale (Fig. 6), IPSL-CM5A and IPSL-CM5A2 were among the best-performing model simulations based on distributional metrics (IQR overlap, KS statistic, and p-values), and IPSL-CM5A also showed good agreement with the reconstructed latitudinal structure of density anomalies (Fig. 3).

Nonetheless, the new performance metrics (RMSE, bias, NSE, and KGE), recommended by reviewer 2 (see answer to reviewer 2), provide a complementary perspective and indicate that iLOVECLIM simulations generally achieve the best overall performance across both LGM and piControl periods. IPSL-CM5A and IPSL-CM5A2 remain among the better-performing CMIP-class models, particularly in terms of distributional agreement, although they are not consistently the top-performing models across all error-based and efficiency metrics.

This will be discussed in more detail in the next version of the paper.

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.
We have clarified this point in the revised manuscript.

We will add these sentences next to Fig.10:

"The very small IQR observed in the Southern Ocean reflects both the limited number of available reconstructions and their relatively homogeneous values. In this region, proxy data

are predominantly located in open-ocean settings, where spatial variability in surface density is lower than in coastal regions, leading to a narrower distribution."

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 will reduce 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 comment. The key question is whether the simulations that best reproduce the mean west-east density of SST gradient in our study also exhibit a realistic level of IOD variability.

We aimed to expand the discussion to better contextualize the reconstructed gradient changes in relation to low-frequency variability of the Indian Ocean Dipole (IOD), following Brierley et al. (2023). But, some limitations prevent a direct comparison.

First, the four simulations that best reproduce the mean west-east gradient in our analysis are not part of the analyse in Brierley et al. (2023). Second, if we had to analyse this four simulations in term of IOD strength, this subset remains too small to provide statistically robust conclusions regarding the relationship between mean-state changes and IOD variability.

More generally, among the 13 simulations analyzed by Brierley et al. (2023), 8 of them overlap with our ensemble, and these simulations do not adequately reproduce the LGM-PI west-east SST or density gradient. This question if the lack of a clear relationship between mean-state changes and IOD variability, for LGM and PI time periods, reported in Brierley et al. (2023) may partly reflect the inclusion of simulations with substantial mean-state biases.

This reflexion will be added in the revised version of the paper.

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

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

We have now added a sentence to acknowledge this result and to place our findings in the context of previous model evaluations. While our analysis confirms that some HadCM3 simulations perform well-particularly in reproducing the west-east density gradient, our results also show that model performance depends on the metric and region considered.

References:

Caley, T., Rieger, N., Werner, M., Waelbroeck, C., Barathieu, H., Happé, T., and Roche, D. M.: Past Ocean surface density from planktonic foraminifera calcite δ18O, Clim. Past, 22, 247–263, https://doi.org/10.5194/cp-22-247-2026, 2026.