

**Using ocean surface paleo-density to evaluate PMIP3 and PMIP4 Last Glacial
Maximum climate simulations.**

Héloïse Barathieu ¹, Thibaut Caley ¹, Masa Kageyama ², Didier Swingedouw ¹, Pascale
Braconnot ²

¹Univ. Bordeaux, CNRS, Bordeaux INP, EPOC, UMR 5805, 33600 Pessac, France

²Laboratoire des Sciences du Climat et de l'Environnement – Institut Pierre-Simon Laplace
(LSCE-IPSL), UMR CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France.

Correspondence to: Héloïse Barathieu (*heloise.barathieu@u-bordeaux.fr*) or Thibaut Caley
thibaut.caley@u-bordeaux.fr)

Abstract

Quantitative reconstruction of ocean surface density during the Last Glacial Maximum (LGM) offers valuable insights into the ability of climate models to simulate past climate conditions, when global temperatures were about 4.5°C to 6°C colder than today. We assess the performance of the LGM climate simulations, as part of the 3rd and 4th phase of the Paleoclimate Modeling Intercomparisons Project, using a recent ocean surface density reconstruction based on the $\delta^{18}\text{O}$ of foraminiferal calcite ($\delta^{18}\text{Oc}$). We consider the differences between the LGM and the preindustrial climates and each period separately, at both global and regional scales. Because surface density reflects the combined effects of temperature and salinity, we also examined sea surface temperature (SST) to better identify the processes underlying model–data differences.

On a global scale, surface density reconstructions generally exhibit greater spatial variability than simulated surface density anomalies (LGM - PI), although part of this difference is reduced when reconstruction uncertainties are taken into account. Model simulations tend to underestimate the magnitude of reconstructed density anomalies and substantial differences between cumulative distribution persist. Part of the mismatch may arise from the uneven spatial distribution of reconstructions, which are mostly located near coastal areas.

Density anomaly (LGM - PI) differences between data and models are largely controlled by sea surface salinity (SSS), with SST contributing to a lesser extent. This influence of SSS is directly linked to the reduction in tropical precipitation during the LGM: model simulations that best match the large-scale density anomalies also simulate the strongest reductions in reconstructed low-latitude precipitation during the LGM, highlighting the key role of hydrological cycle changes in shaping surface density.

All simulations capture key features of the reconstructed surface densities when the LGM and PI periods are analysed separately at the global scale, with Taylor diagrams and complementary performance metrics indicating moderate to good overall agreement despite differences among simulations.

Regional analyses show that most simulations reproduce reconstructed Indian Ocean surface density reasonably well, **although model simulations performance is systematically lower in the North Indian Ocean than in the South Indian Ocean**. Focusing on the Indo-Pacific Warm Pool, proxy reconstructions indicate a weakened West-East tropical Indian surface density gradient during the LGM, but only 7 out of 14 model simulations (50%) reproduce this feature.

These results highlight the need to improve and better constrain regional hydrological cycle changes in models, as improving their representation is crucial to reduce uncertainties in both paleoclimate simulations and future climate projections.

1. Introduction

Past surface seawater density is a key property for studying ocean dynamics, as it reflects the combined influence of surface temperature and salinity and is directly linked to circulation changes through geostrophic balance. In this study, we focus specifically on seawater surface density, providing a novel perspective in model–data comparisons for the Last Glacial Maximum (LGM), a variable that has not been explored in previous assessments.

To simulate future climate change, scientists rely on coupled general circulation models (GCMs). However, these models differ in their representation of Sea Surface Temperature (SST), Sea Surface Salinity (SSS), as well as in the processes that control density. Differences in changes of SST and SSS in the future is therefore leading to large uncertainties in the simulation of future ocean dynamics (Flato et al., 2014 – IPCC AR5; Eyring et al., 2021 – IPCC AR6).

Since climate models are developed based on present-day conditions and used to project future climates that may be very different from the present one, it is important to also evaluate them with reconstructed very different climate conditions from the past, to gain confidence in their projections. One way to test the response of these models to various external forcings is to use paleoclimate simulations, which provide an independent evaluation of model performance (Harrison et al., 2014; Kageyama et al., 2024) against available reconstructions. This allows us to benchmark models using evidence from past climates, which is essential for strengthening confidence in their future projections. The Paleoclimate Model Intercomparison Project (PMIP; Joussaume and Taylor, 1995; Kageyama et al., 2018) tests the ability of models to simulate paleoclimate reconstructions. Currently in its fourth phase (PMIP4), with a fifth in preparation, PMIP plays a critical role in evaluating how well models reproduce past climates.

One of the PMIP reference periods is the Last Glacial Maximum, which occurred between 19,000 and 23,000 years ago, when the ice sheets reached their maximum global volume (Mix et al., 2001). During this period, the climate was markedly different from pre-industrial conditions, with significantly colder temperatures (from -4.5 ± 0.9 °C according to Annan et al. 2022 to -6.1 ± 0.4 °C according to Tierney et al. 2020) and altered hydrological cycles, making it an interesting benchmark period for evaluating climate models (MARGO project, 2009; Braconnot et al., 2012).

Evaluating numerical climate models using a data-model comparison allows us to assess their robustness in simulating key variables such as sea surface temperature (SST) and precipitation (Brierley et al., 2023). Recent intercomparison studies focusing on SST reconstructions at the LGM (Tierney et al., 2020; Kageyama et al., 2021) show that models generally capture large-scale cooling patterns but still often exhibit regional biases and differences between PMIP3 and PMIP4 simulations.

To enable quantitative evaluation of ocean surface density, a new method has been developed to reconstruct annual seawater surface densities in the past (Caley et al., 2026), providing a novel tool for model assessment. However, until now, quantitative evaluations of

surface density remain unexplored, which limits our understanding of how well models capture the combined effects of temperature and salinity on ocean circulation.

In this study, we make use of this new surface density reconstruction to evaluate PMIP3 and PMIP4 simulations in terms of annual ocean surface density, both on a global scale (excluding the Nordic Seas region, Caley et al., 2026) and regionally. We consider simulations of the Last Glacial Maximum (LGM) and the pre-industrial period (piControl). We focus on evaluating model performance in simulating surface density, and where there are large discrepancies between models and past reconstructions, we further investigate SST and, in combination with density, qualitative changes in SSS. This approach provides a complementary perspective to previous studies, offering new insights into the coupled role of temperature and salinity in shaping ocean surface density and allowing for a more integrated evaluation of model performance.

Our analysis is structured as follows: we first assess model simulations against past reconstructions at the global scale in Sect. 3. We then examine regional differences in Sect. 4, identifying areas where model simulations perform better or worse, with a particular focus on the Indian Ocean as a case study of regional variability.

2. Material and methods

2.1. Climate reconstructions

2.1.1. Surface ocean density

To evaluate model simulations on ocean surface density, we use the quantitative past density reconstruction dataset proposed by Caley et al. (2026). They developed a new Bayesian calibration model to calculate the annual surface ocean density using the $\delta^{18}\text{O}_\text{c}$ measurements of several foraminiferal species. Briefly, this probabilistic approach explicitly accounts for inter-species differences and calibration uncertainties, allowing quantitative density reconstructions. New and published $\delta^{18}\text{O}_\text{c}$ 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 $\delta^{18}\text{O}_\text{c}$ 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 $> 40^\circ\text{N}$ 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/m^3 . Throughout this work, values are expressed as anomalies relative to 1000 kg/m^3 .

Concerning the instrumental observations, we used the version 4.2.2 (analyses.g10, downloaded in 2024) of the EN dataset from the Met Office Hadley Centre (Good et al., 2013), commonly referred to as EN4. This dataset provides quality-controlled ocean temperature and

salinity profiles globally, as well as monthly gridded fields derived from objective analyses, covering the period from 1900 to 2022. EN4 is a compilation of temperature and, when available, salinity measurements from various ocean data sources. Surface seawater density, which is a non-linear function of temperature and salinity, was calculated from EN4 temperature and salinity fields using the GSW (Gibbs Seawater) formulation implemented in the gsw Python package (Roquet et al., 2015).

2.1.2. Sea surface temperature

Since sea surface density and SST are linked, we also performed a data/model comparison in terms of surface temperature. This provides an additional way to investigate the drivers of surface density changes. To do this, we used two previously published SST databases (MARGO Project, 2009; Tierney et al., 2020). These two databases were not combined, as the Tierney dataset includes some MARGO data but with more recent calibrations. Notably, Tierney et al. (2020) also recalibrated all age models using the Marine13 radiocarbon calibration curve and the BACON age modelling software, ensuring a better chronological consistency across records. The temporal periods differ slightly: Tierney’s data refer to the LGM and the Late Holocene (LH), while the MARGO data include LGM and “pre-industrial” measurements from the WOA1998 dataset (NODC, Silver Springs, 1998). Following Tierney et al. (2020), we make here the approximation that the Late Holocene is considered representative of the pre-industrial climate state.

The MARGO database (MARGO Project, 2009) contains 821 SST reconstructions based on a diverse range of proxies, including Mg/Ca ratios, $U_{k'37}$ indices, radiolarians, diatoms, foraminiferal transfer functions, and the tetraether index TEX_{86} . In contrast, Tierney’s dataset (Tierney et al., 2020) comprises 244 SST records derived exclusively from Mg/Ca, TEX_{86} , and $U_{k'37}$ proxies. We do not use the assimilated SST product developed by Tierney et al. (2020), but only the raw SST proxy database published in association with their study. This selection reflects a deliberate choice by Tierney et al. (2020) to exclude assemblage-based proxies such as foraminiferal transfer functions, due to concerns over “no-analogue” assemblages and the lack of Bayesian calibration models, which are central to their probabilistic framework. SST values inferred from $\delta^{18}O_c$ were also excluded from our analysis, as they were already incorporated into our density reconstructions. Finally, both SST and density datasets were re-gridded onto a common $1^\circ \times 1^\circ$ spatial grid, matching the reference grid to which the model simulations were also re-gridded, allowing for a direct comparison.

2.2 Climate model simulations

For this study, we used LGM and pre-industrial (hereafter piControl) climate model simulations from a total of sixteen simulations, including seven from PMIP3 (Braconnot et al., 2012) and nine from PMIP4 (Kageyama et al., 2018) (see Table 1). Two more simulations were excluded due to inconsistencies in salinity data (e.g. unit or formatting issues), making them unsuitable for analysis. Some model simulations share the same piControl but differ by their imposed ice-sheet reconstructions for LGM (e.g., HadCM3-ICE6GC vs. HadCM3-GLAC1D and iLOVECLIM1-1-1-GLAC1D vs iLOVECLIM-1-1-1-ICE-6G-C). We tested whether simulations from the same model (within a PMIP phase or across PMIP3/PMIP4 versions) were too similar to

each other. Our analysis showed that all simulations differed in at least one basin and for at least one of the variables (SSS, SST, density). Based on this, we retained all simulations.

The pre-industrial control simulation uses the constant boundary conditions established for 1850 CE (Eyring et al., 2016). It aims to produce a stable quasi-equilibrium climate under 1850 conditions, characterised by the annual cycle (mean and seasonality) and internal variability arising from interactions between Earth system components. This simulation serves as a baseline from which changes in all other experiments are calculated.

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 models, 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/m³ density correction to the reconstructions, following Caley et al. (2026).

We analysed annual mean SSS and SST from these simulations. All outputs were regridded to a common $1^\circ \times 1^\circ$ grid, and monthly data were averaged to obtain climatological annual means. These means were computed over the full duration of each simulation, which ranges from 100 to 1100 years for the piControl experiments and up to 500 years for the LGM experiments.

The variables used here are salinity (so) and temperature (thetao). To study surface density, we used salinity and temperature from the first ocean layer. Seawater density was calculated using the gsw (Gibbs Seawater) Python package, which is based on the TEOS-10 thermodynamic framework and provides thermodynamically consistent equations for seawater properties. The gsw package requires Absolute Salinity (SA) and Conservative Temperature (CT) as inputs. CT is a more accurate and thermodynamically sound analog of Potential Temperature (θ), and SA is derived from Practical Salinity (SP), which is unitless and not directly usable in thermodynamic calculations. Since the model simulations originally provided SP and θ , we first converted them to SA and CT using the functions SA_from_SP and CT_from_pt, respectively. Finally, in-situ seawater density was computed with gsw.rho(SA, CT, p), using a computationally efficient expression for specific volume as a function of SA, CT, and pressure (Roquet et al., 2015).

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 r1i1p1	0.672	Sénési et al. (2014 - LGM and PI) Voldoire et al. (2013)
MIROC-ESM	MIROC-ESM	PMIP3-CMIP5 r1i1p1	1.002	JAMSTEC et al. (2015 - LGM and PI) Sueyoshi et al. (2013)
MPI-ESM-P	MPI-ESM-P	PMIP3-CMIP5 r1i1p1	0.924	Jungclaus et al. (2014 - LGM and PI) Adloff et al. (2018)
MRI-CGCM3	MRI-CGCM3	PMIP3-CMIP5 r1i1p1	1.055	Yukimoto et al. (2012) Yukimoto et al. (2015 - LGM and PI)
GISS-E2-R	GISS-E2-R	PMIP3-CMIP5 r1i1p150	1.269	NASA/GISS (2014 - LGM and PI) Schmidt et al. (2014) Ullman et al. (2014)
IPSL-CM5A-LR	IPSL-CM5A	PMIP3-CMIP5 r1i1p1	1.357	Kageyama et al. (2016 - LGM) Caubel et al. (2016 - PI)
CCSM4	NCAR-CCSM4	PMIP3-CMIP5 r1i1p1	0.998	Gent et al. (2011) Brady et al. (2013)
<i>HadCM3B-M2.1aD</i>	<i>HadCM3-PMIP3</i>	<i>PMIP4-CMIP6 PMIP3 ice sheets</i>	1.172	<i>Izumi et al. (2023)</i>
<i>HadCM3B-M2.1aD</i>	<i>HadCM3-GLAC</i>	<i>PMIP4-CMIP6 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 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 r1i1p1f1</i>	0.919	<i>Mauritsen et al. (2019)</i>
<i>MIROC-ES2L</i>	<i>MIROC-ES2L</i>	<i>PMIP4-CMIP6 r1i1p1f2</i>	1.115	<i>Ohgaito et al. (2019 - LGM) 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 GLAC-1D ice sheet</i>	1.450	<i>Lhardy et al. (2021) Bouttes et al. (2023)</i>
<i>iLOVECLIM1-1-1-ICE-6G-C</i>	<i>iLOVECLIM-ICE-6G-C</i>	<i>PMIP4-CMIP6 ICE-6G-C ice sheet</i>	1.641	<i>Lhardy et al. (2021) Bouttes et al. (2023)</i>
<i>CESM1-2</i>	<i>CESM1.2</i>	<i>PMIP4-CMIP6</i>	1.215	<i>Hurrell et al. (2013) Tierney et al. (2020)</i>

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/m³). References correspond to the model description papers and/or ESGF simulation references when available.

2.3 Statistical analysis:

Model–data comparisons were performed on a $1^\circ \times 1^\circ$ grid, restricted to locations where proxy reconstructions provide valid values. Model anomalies were extracted exactly at the proxy sites to ensure strict spatial consistency. 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.

Distributional agreement between models and reconstructions was evaluated using three complementary criteria. First, interquartile range (IQR) overlap was assessed to check distributional consistency while limiting sensitivity to extreme outliers. Second, the Kolmogorov–Smirnov (KS) statistic, which quantifies the maximal difference between cumulative distributions. 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. Third, a two-sample KS test p -value < 0.05 indicates a statistically significant difference between distributions: a value above 0.05 indicates that the null hypothesis of identical distributions cannot be rejected, while $p < 0.05$ indicates a significant difference. For visualization, kernel density estimates (KDEs) provided smoothed, continuous representations of the distributions, highlighting central tendencies and the most frequent values.

For the global density anomaly distribution analysis (Sect. 3.2), reconstruction uncertainty was additionally propagated using a Monte Carlo approach (10,000 iterations). For each reconstruction site, density anomalies were randomly sampled within their asymmetric 95 % confidence intervals derived from the Bayesian calibration model (Caley et al., 2026). The reconstructed median, interquartile range (IQR), and KS statistics were then recalculated for each realization. This procedure allowed us to evaluate how reconstruction uncertainty affects the distributional comparison between models and data.

For linear relationships, regression analyses were performed separately for PI and LGM periods. The coefficient of determination (R^2) quantifies the proportion of variance in reconstructions explained by the models, while the slope measures the amplitude of the model response relative to observations. To account for uncertainties in the reconstructions, a Monte Carlo procedure (10,000 iterations) added Gaussian noise to the observations, derived from the 95% confidence intervals. Distributions of R^2 and slope were then analyzed, and reported values correspond to mean $\pm$ standard deviation.

To further compare data and model simulations we introduce other metrics to reduce overreliance on regression and correlation. 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,

284 providing an integrated measure of model performance, with values close to 1 indicating
285 better agreement.

286 This framework provides a coherent and rigorous assessment of model-data agreement by
287 combining distribution-based comparisons with analyses of linear relationships. It integrates
288 multiple complementary metrics to evaluate performance, and accounts for reconstruction
289 uncertainties through Monte Carlo approaches. Together, these methods offer a
290 comprehensive evaluation of how well model simulations reproduce observed fields, both
291 globally and across individual ocean basins.

292

293 *2.4 Testing spatial representativeness with a pseudo-proxy approach*

294

295 The spatial distribution of reconstructions is uneven, with a clear concentration of data near
296 coastal areas (Fig. 1, Fig. S1).

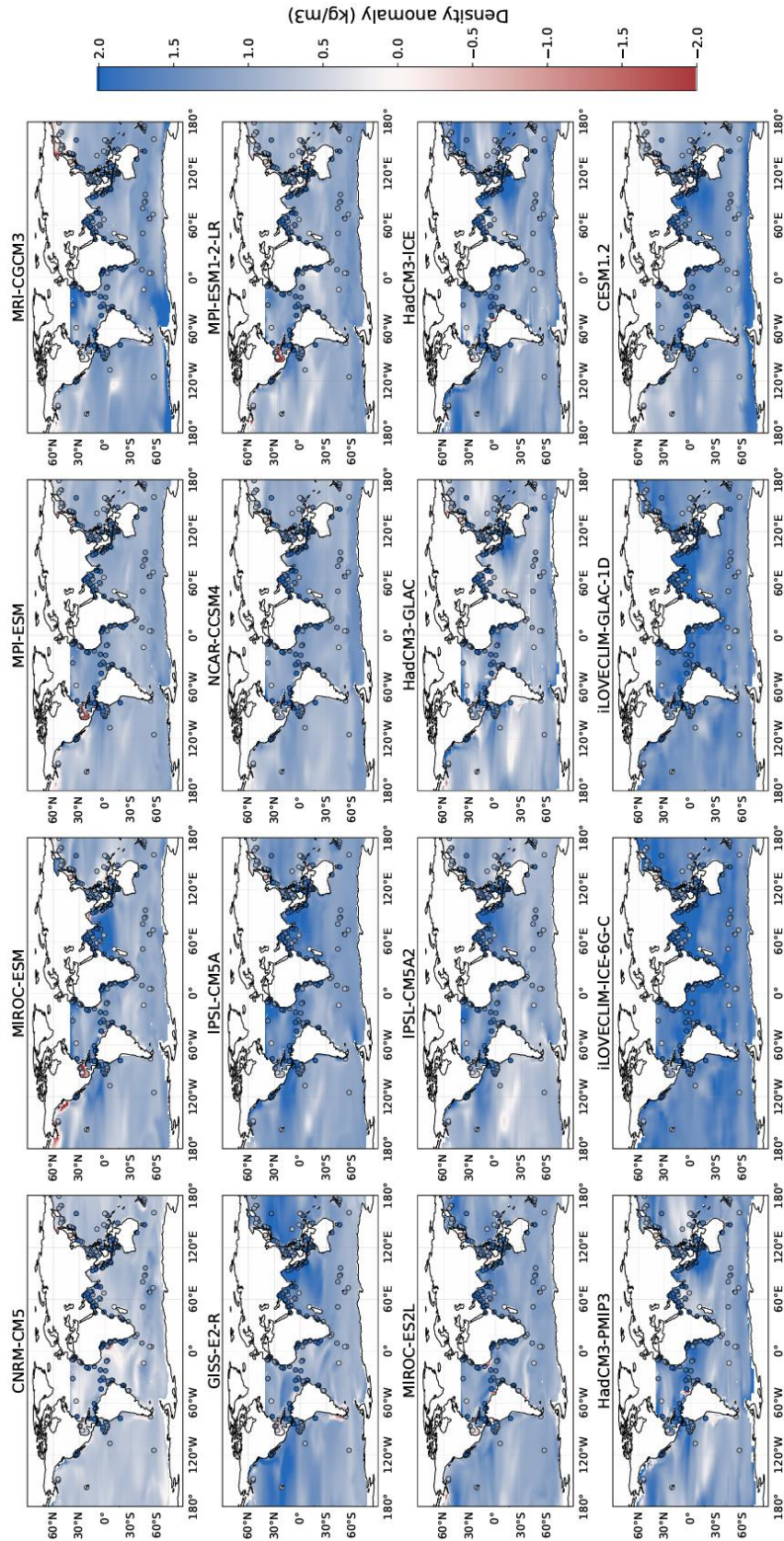


Figure 1: Absolute surface density (kg/m^3) anomaly map (LGM - piControl). The dots represent the surface density anomaly database reconstructions (Caley et al., 2026) and the background map is the anomaly mean for each model simulations in the study.

To assess whether the proxy data locations are representative of broader basin-scale conditions, a “pseudo-proxy” approach (Ayache et al., 2018) was performed, in order to compare the mean local values with basin-wide means derived from the models (Fig. 2, Fig. S2). 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. This analysis evaluates the spatial representativeness of proxy locations at the basin scale. However, it does not address differences between coastal and open-ocean environments, since climate models do not explicitly represent coastal processes and therefore cannot accurately simulate nearshore dynamics. Details on the basin definition and spatial masks used for this analysis are provided in Appendix A.

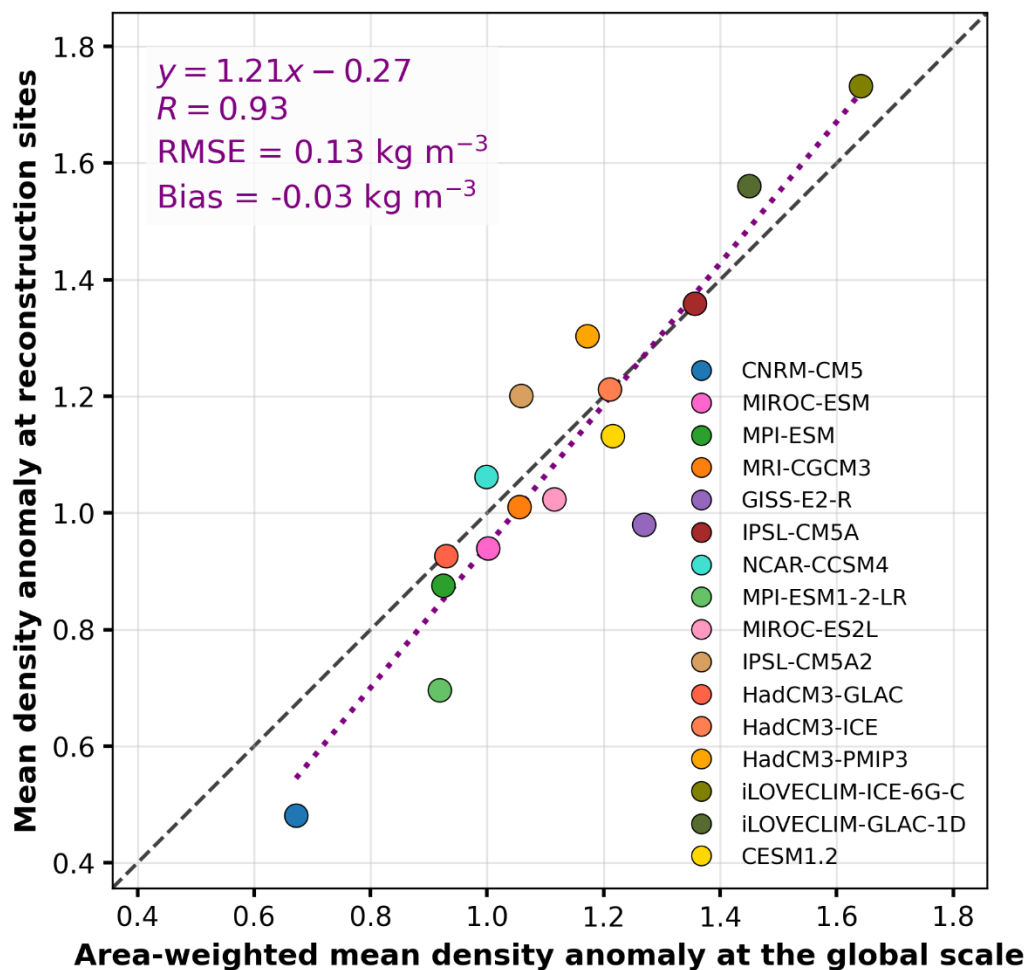


Figure 2: Pseudo-proxy test for global ocean surface density. Comparison between the mean density anomalies (kg/m³) averaged across all model grid points (x-axis) and the mean density anomalies (kg/m³) averaged over proxy reconstruction sites (y-axis). The purple regression line and R represent the fit across all model simulations (PMIP3 + PMIP4). Results for individual basins are provided in Supplementary Fig.S2.

For this purpose, we compare the average of density simulated by PMIP3 and PMIP4 model simulations over a given basin with the average density only at locations where proxy data are

available (Fig. 2 and Fig.S2). A near-linear relationship is found at the global scale (Fig. 2) with the coefficient of determination (R) at 0.93 and across most basins (Fig. S2), when compiling PMIP3 and PMIP4. These regressions indicate how well the mean over proxy sites reproduces the true basin-wide mean in the model simulations. The slopes being close to one and the high R^2 values (p -value < 0.05) indicate that the mean anomalies at the proxy sites capture the basin-wide means simulated by the models reasonably well. This suggests that, within the models, the uneven proxy distribution does not strongly bias the large-scale signal.

3. Evaluate model simulations at the global scale

Before investigating regional features, we first assess the large-scale behaviour of model simulations. A global-scale evaluation allows us to assess the overall ability of PMIP3 and PMIP4 simulations to reproduce the reconstructed large-scale signal of surface density changes between the LGM and the pre-industrial period. Evaluating models at this integrated scale also helps reduce the influence of local reconstruction uncertainties and highlights the dominant climatic drivers of density variations, such as global temperature and hydrological cycle changes.

3.1. Sea surface density anomaly (LGM-PI)

We first analyse absolute surface density anomalies between the LGM and the Pre-Industrial period (LGM-PI), in order to reduce the impact of potential systematic model-specific biases that may persist across time periods. All models simulate positive anomalies, i.e. they agree on the sign of the change. However, both the spatial patterns and the amplitude of the simulated anomalies vary considerably from one model simulation to another (Fig. 1). When averaged over space, simulated mean density anomalies typically range from 0.7 to 1.6, with an average value close to 1.1. In contrast, the proxy-based reconstructions exhibit a slightly higher mean, approximately 1.6.

When zonally averaged, the density anomaly (LGM-PI) in the observations (Fig. 3, grey dots) is stronger in the low latitudes than in the mid-latitudes, as already discussed in Caley et al., 2026. The observed increase in density during the LGM, both in the simulations and in the reconstructions, is consistent with the SST cooling (MARGO project, 2009; Tierney et al., 2020) and with a weaker hydrological cycle at low latitudes, in which precipitation decreased more than evaporation. This reduction in precipitation leads to saltier and denser surface waters (Kageyama et al., 2021), as already discussed in Caley et al., 2026.

Some model simulations fail to reproduce the full latitudinal structure, such as CNRM-CM5 and MPI-ESM-P, while others do not capture the shape but match the density anomaly well at low latitudes, for instance iLOVECLIM-ICE-6G-C and iLOVECLIM-GLAC-1D. Some model simulations, such as IPSL-CM5A and HadCM3-PMIP3, are in good agreement with the data across all latitudes, especially between 0 and 40°N, considering the uncertainties on the

reconstructions (Fig. 3). The same type of zonally averaged analysis was performed for SSS (Fig. B2) and for SSTs, the corresponding results are shown in Figs. B3 and B4.

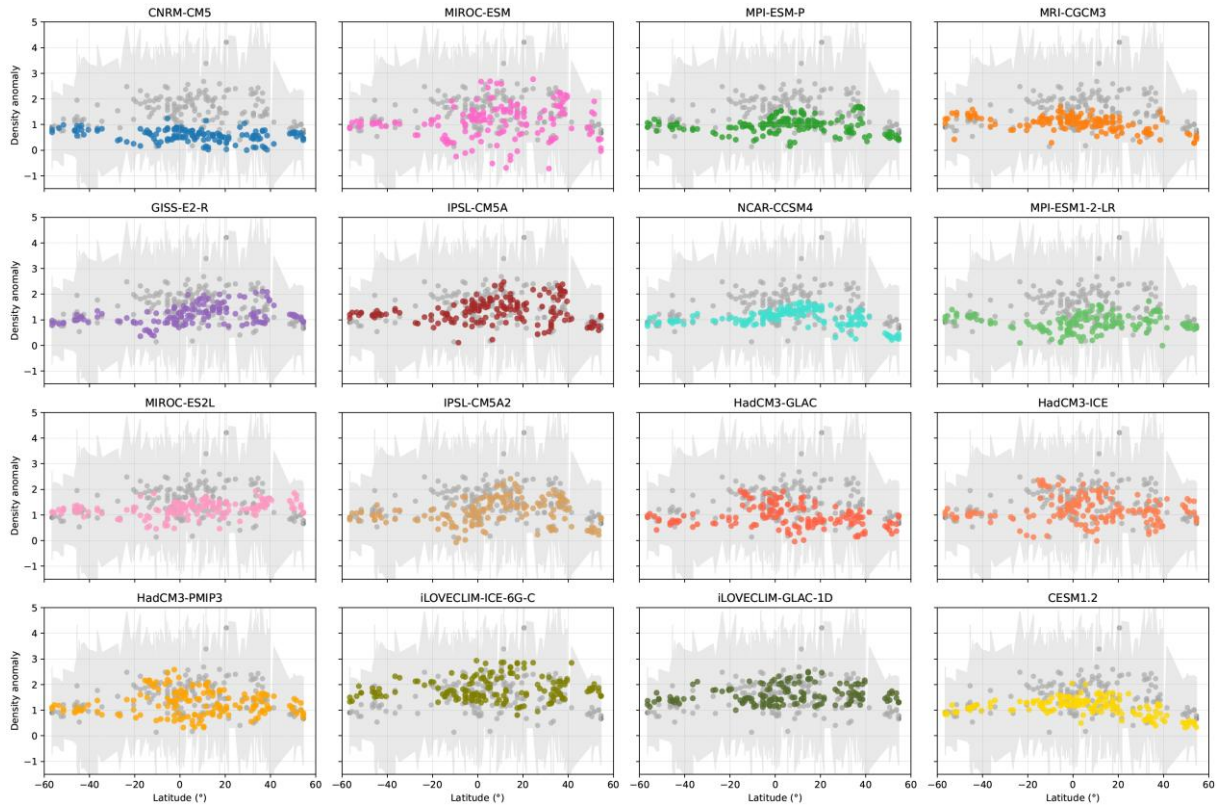


Figure 3: Density anomaly (kg/m^3) as a function of latitude for each model simulation (colored dots) compared with the observational data (grey dots) and the 95% confidence interval (grey shading). Model outliers, identified using the interquartile range (IQR) method (values outside $1.5 \times \text{IQR}$), 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.

We next investigated the physical drivers of the density differences between reconstructions and simulations by decomposing the total density anomaly into temperature and salinity components. Model outputs and proxy datasets were collocated at common sampling points without additional interpolation, ensuring a strict one-to-one correspondence between density and SST observations. This procedure yielded 80 common points for the MARGO database (MARGO Project, 2009) and 92 points for the SST dataset from Tierney et al. (2020). No outlier filtering was applied.

Before computing the salinity contribution, the global density related to sea-level-induced salinity increase (-1 g/kg in models, -0.77 kg/m^3 in reconstructions; Caley et al., 2026) was removed to isolate density changes linked to hydrographic changes in SST and SSS. Because this correction is applied consistently to both reconstructions and model outputs, and since the analysis is based on model–data differences, it has no effect on the results presented here.

To quantify the relative influence of temperature and salinity on the model–data density anomaly differences, we decomposed the total density anomaly difference (model simulations minus reconstructions) into thermal and haline components using a Shapley decomposition. This approach accounts for the non-linearity of the seawater equation of state by averaging the contributions obtained from the two possible attribution sequences

(temperature-first and salinity-first). The method provides an order-independent partitioning of the total density anomaly while ensuring that the sum of the thermal and haline contributions exactly reproduces the total density difference.

Figure 4 shows the relative contributions (%) of SST and salinity to the density anomaly difference for each model simulation. The relative importance of both components varies among simulations and between SST reconstruction datasets. For the MARGO-based estimates, SST and salinity contributions are often of comparable magnitude, whereas the Tierney-based estimates generally show a larger contribution from salinity. Nevertheless, SST effects alone cannot explain the differences in data-model density, demonstrating that hydrological changes substantially contribute to the differences in model-data density anomalies. It is also important to note that Figure 4 presents basin-averaged contributions and therefore smooths part of the latitudinal variability observed in the density anomalies. The zonally averaged distributions shown in Fig. B2 reveal that this spatial density (Figure 3) variability is primarily controlled by salinity anomalies, whereas SST anomalies exhibit much weaker latitudinal variations across model simulations.

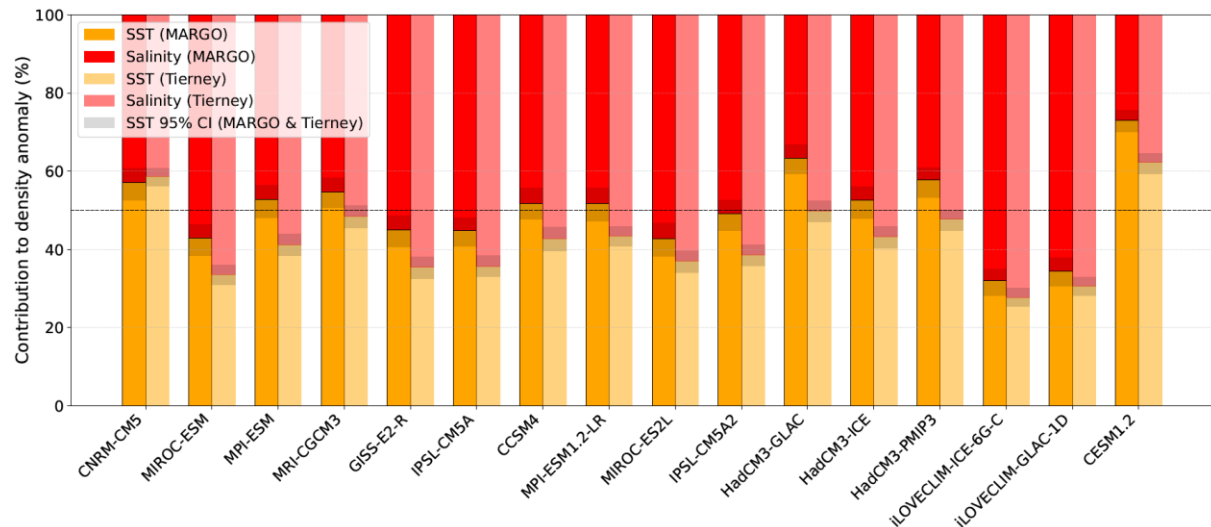


Figure 4: Relative contributions (%) of SST (orange) and salinity (red) to the model–data density anomaly differences (LGM–PI). Contributions were estimated using a Shapley decomposition, which provides an order-independent partitioning of density changes into thermal and haline components while accounting for the non-linearity of the seawater equation of state. Dark colours correspond to SST reconstructions from the MARGO Project (2009), whereas lighter colours correspond to Tierney et al. (2020). Shaded envelopes indicate the 95% confidence intervals obtained by bootstrap propagation of SST reconstruction uncertainties.

Given that the hydrological cycle influences surface salinity and that salinity anomalies strongly influence the difference in data/model density anomalies, we explored the link between density anomalies and large-scale precipitation changes. Using the PMIP3 and PMIP4 ensembles, Kageyama et al. (2021) showed that nearly all models simulate substantial decreases in precipitation in high-rainfall regions during the LGM, particularly across the tropics and monsoon zones, although the magnitude varies between model simulations. Model simulations that best reproduce the density anomalies inferred from proxy data also tend to exhibit the largest reductions in tropical precipitation. In contrast, models with smaller precipitation decreases fail to reproduce the observed density structure. Figure 5 illustrates the relationship between density anomalies and mean annual precipitation anomalies for

both the tropics (30°S–30°N, circle) and the global ocean (square). A clear linear relationship emerges between density anomalies and precipitation anomalies, with $R^2 = 0.67$ in the tropics and $R^2 = 0.64$ globally, highlighting the robustness of this connection (p -value < 0.05). This analysis emphasizes that accurately representing low-latitude hydrological feedbacks is critical for capturing the full magnitude of glacial ocean density changes.

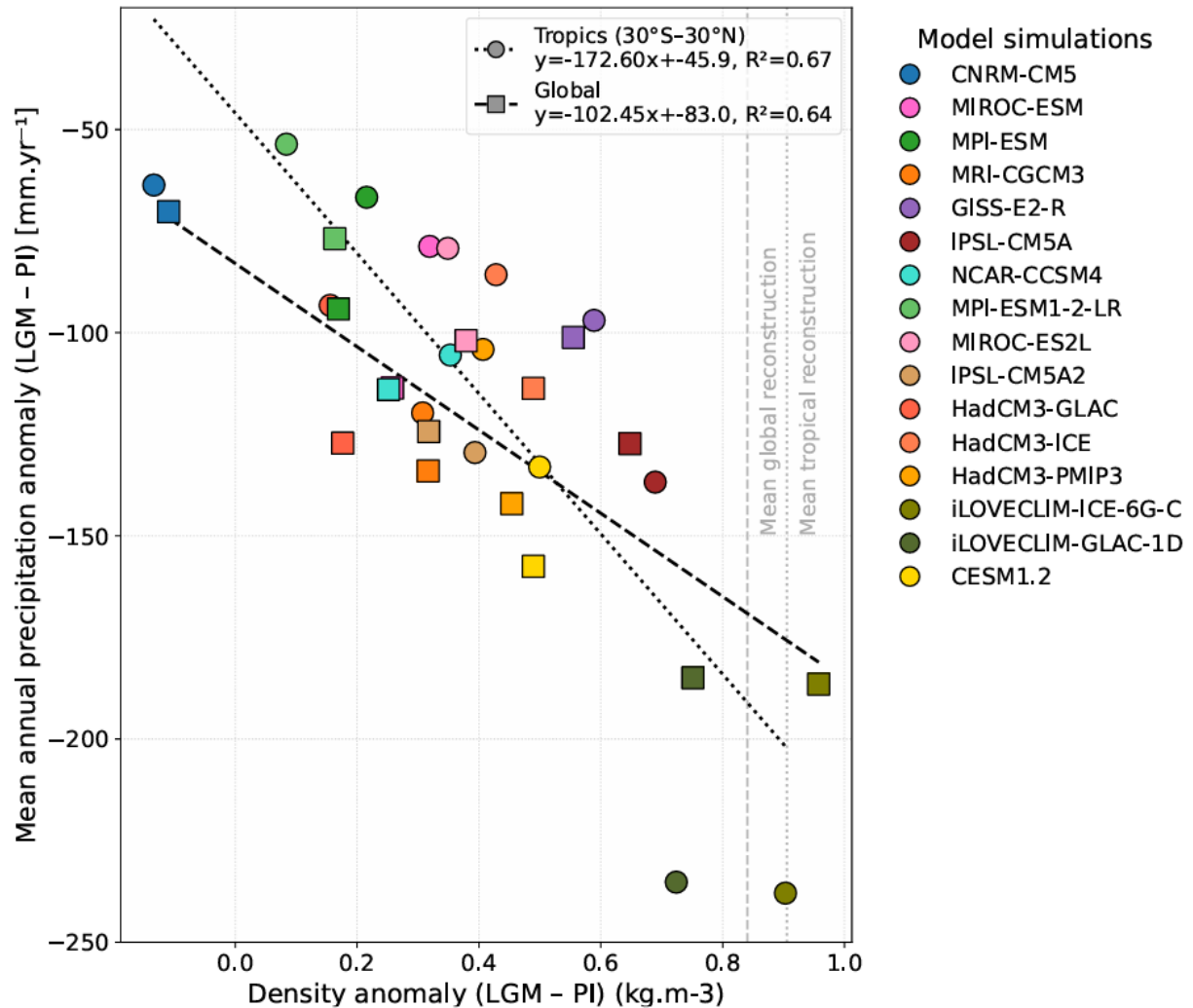


Figure 5: Mean annual precipitation anomalies (LGM – PI, mm.yr^{-1} , over land and ocean) as a function of seawater density anomalies corrected for mean ocean density changes (kg/m^3) relative to sea level-induced salinity increase at LGM. Scatter points show the relationship for the tropics (30°S–30°N, 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.

In summary, our results indicate that SST effects alone cannot explain the density anomaly differences between reconstructions and simulations. Instead, salinity differences account for most of the model–data anomaly density discrepancy (Fig. 4) and are directly linked to reductions in tropical precipitation. Models that simulate stronger tropical precipitation decreases reproduce the observed LGM surface density anomalies more accurately,

emphasizing the importance of representing low-latitude hydrological feedbacks to capture the full magnitude of glacial ocean density changes.

3.2. Comparison of global distribution of surface density anomalies

Before analysing regional contrasts, we first evaluate model performances considering the global distribution of absolute surface density anomalies (LGM-PI) by aggregating data from all selected ocean basins (Fig. 6 & Fig. 7, Fig. C1).

Model-data agreement was evaluated using three complementary statistical criteria (see Sect. 2.2): a Monte Carlo-derived p-value > 0.05 from a two-sample Kolmogorov-Smirnov (KS) test, a Monte Carlo-derived KS statistic below 0.13 and overlapping interquartile ranges (IQRs) between data and model distributions, indicating consistent variability. Together, these criteria evaluate whether models reproduce not only the central value but also the overall shape of the reconstructed distribution.

Figure 6 visually illustrates these three criteria through the coloured flags, which show whether the Monte Carlo-derived KS statistic, p-value, and IQR criteria are satisfied for each model simulation. Figure 7 complements this by showing only the IQR and median values of the model and reconstruction distributions, including a Monte Carlo propagated reconstruction uncertainty, allowing a direct comparison of central tendency and spread.

Reconstruction-based anomalies have a median value of 1.19, whereas the values obtained from model simulations vary widely, ranging from 0.6 to 1.6 (Fig. 6 and Fig. C1). Interquartile ranges (IQRs) support this divergence. Using the reconstruction values without considering reconstructions uncertainties, the reconstructions exhibit a broader spread (IQR ~0.75), while models show smaller variability (IQR between 0.27 and 0.98). However, when reconstruction uncertainties are propagated through the Monte Carlo procedure, the reconstructed IQR increases substantially, reflecting the uncertainty associated with the density calibration. As a result, all model simulations fall within the Monte Carlo-derived reconstruction IQR envelope (Fig. 6 and Fig.7), indicating that differences in distributional spread alone are no longer sufficient to discriminate between simulations. Once reconstruction uncertainty is propagated, the IQR criterion is satisfied by all simulations, leaving the Monte Carlo-derived KS statistic and associated p-value as the main metrics discriminating among model simulations.

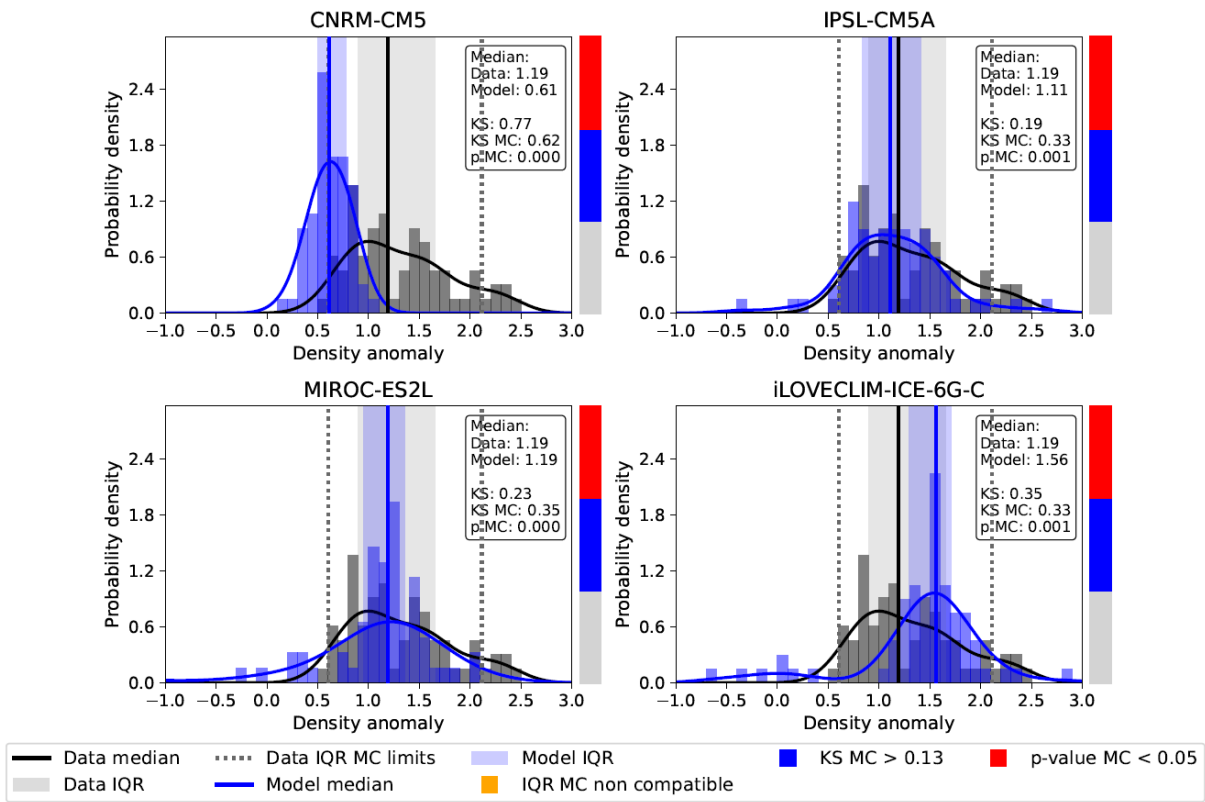


Figure 6: Distribution histograms of surface density anomalies (LGM-PI, kg/m3) for an example of four model simulations. Density reconstructions are shown in black, model simulations in blue. Kernel Density Estimates (KDEs) illustrate the central tendency and overall shape of the distributions. Vertical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

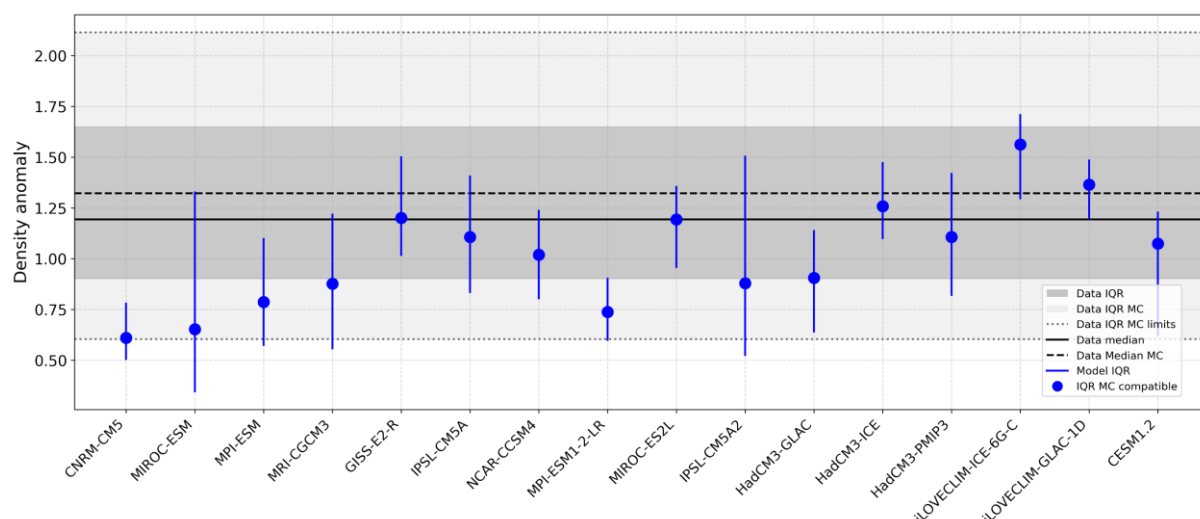


Figure 7: The dark grey band indicates the reconstructed IQR without considering uncertainties, while the light grey band and grey dotted lines indicate the Monte Carlo propagated reconstruction IQR after accounting for reconstruction uncertainty. The solid black line marks the reconstructed median and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

As shown in Fig. 6 and Fig. 7, none of the simulations satisfy all three criteria simultaneously. Although all simulations satisfy the IQR criterion, all Monte Carlo-derived KS statistics remain above the prescribed threshold (0.13) and all Monte Carlo-derived p-values remain below 0.05. This indicates that, although the observed range of variability can be reconciled once reconstruction uncertainty is considered, important discrepancies persist in the overall shape of the modelled and reconstructed distributions. Consequently, agreement between reconstructed and simulated global distributions remains limited.

Because density anomalies depend on both temperature and salinity, we also compared the distribution of SST anomalies between models and reconstructions using the MARGO (MARGO project, 2009) and Tierney et al. (2020) datasets (Fig. S3 and S4). Using the MARGO (MARGO project, 2009) reconstructions, 13 out of 16 simulations (81%) satisfy the Monte Carlo propagated IQR criterion, indicating that model and proxy interquartile ranges remain largely consistent after accounting for reconstruction uncertainty. However, all Monte Carlo-derived KS statistics remain above the prescribed threshold and all Monte Carlo-derived p-values remain below 0.05, indicating that although the overall spread of modelled and reconstructed SST anomalies is comparable, their overall distribution remain statistically different.

Similarly, for the Tierney et al. (2020) dataset, 14 out of 16 simulations (87.5%) satisfy the Monte Carlo propagated IQR criterion and 5 out of 16 simulations (~30%) satisfy the MC p-value limit. For the MARGO comparison, all Monte Carlo-derived KS statistics exceed the prescribed threshold and all Monte Carlo-derived p-values remain below 0.05, indicating that none of the simulations reproduce the overall distribution of reconstructed SST anomalies despite generally consistent interquartile ranges.

Finally, part of the mismatch may arise from the uneven spatial distribution of reconstructions, which are mostly located near coastal areas (Fig. 1). These coastal regions are particularly

complex to simulate due to influences such as continental runoff and oceanic upwelling, which are often poorly captured by global climate models. Consequently, kernel density estimates derived from reconstructions appear flatter, suggesting greater variability not captured by the models. Additionally, reconstructions in some key upwelling zones remain problematic, as highlighted by Caley et al. (2026), further complicating the comparison and evaluation of model outputs against data.

3.3. Global evaluation of surface density: models vs. reconstructions (PI & LGM)

One limitation of working with anomalies is that any observed difference between data and model simulations in absolute surface density cannot be directly linked to either the LGM or the PI baseline. To address this, we analysed the two periods separately, the LGM and the PI, as shown in Fig. 8 and Fig. 9.

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.

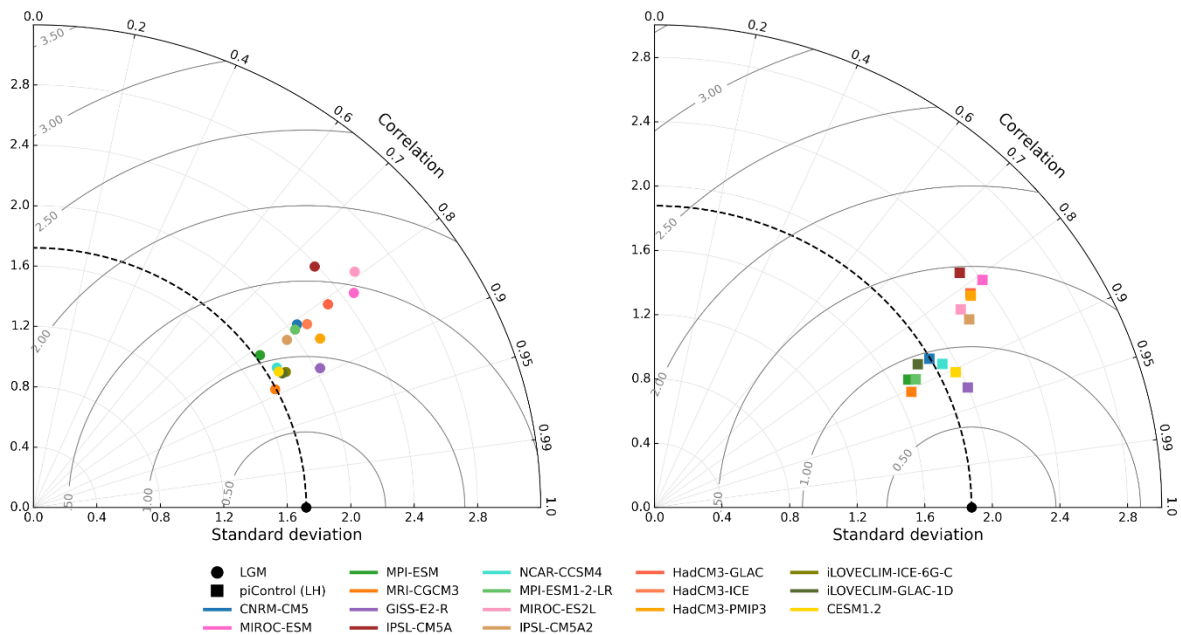


Figure 8: Taylor diagrams comparing model simulations performance for LGM (left) and piControl (LH, right) ocean surface density reconstructions. All grid cells are equally weighted in these diagrams. Each point represents a climate model simulation. Circles correspond to the Last Glacial Maximum (LGM), 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.

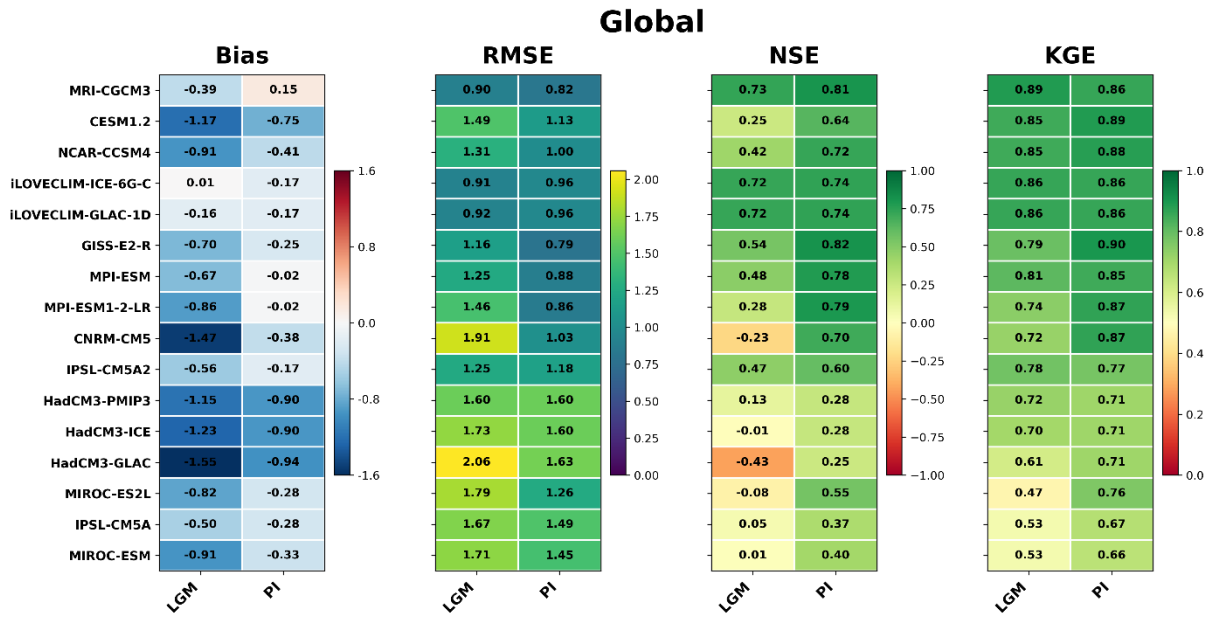


Figure 9: Global model simulations performance for sea surface density during the LGM and PI periods. Heatmaps show four complementary metrics calculated from model–data comparisons at reconstruction grid cells common to all simulations: bias, root mean square error (RMSE), Nash–Sutcliffe efficiency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

Overall, the Taylor diagrams (Fig. 8) and the complementary performance metrics (Fig. 9) provide a consistent assessment of model simulations skill. During the LGM, the two iLOVECLIM simulations remain among the best-performing models, combining the lowest RMSE values (0.90–0.91 kg m⁻³), high NSE values (0.72–0.73), and KGE values of 0.86–0.89, while also exhibiting excellent agreement in the Taylor diagram. MRI-CGCM3 performs similarly well, with the highest KGE (0.89), a low RMSE (0.92 kg m⁻³), and a relatively small bias (–0.39 kg m⁻³). IPSL-CM5A2 and MPI-ESM also rank among the strongest simulations, indicating that several model simulations reproduce both the spatial variability and the magnitude of reconstructed surface densities with good overall skill. Rather than identifying a single best-performing simulation, the different metrics consistently highlight a small group of high-performing model simulations.

In contrast, HadCM3-GLAC, HadCM3-ICE, HadCM3-PMIP3 and MIROC-ESM consistently rank among the **least skilful** simulations. These models combine relatively large RMSE values, pronounced negative biases, lower correlations in the Taylor diagrams, and reduced NSE and KGE values, indicating persistent difficulties in reproducing the mean state, the amplitude of spatial variability, and the spatial pattern of reconstructed surface densities. These differences are also evident in the regression plots (Fig. D1), where the best-performing models display data points clustered close to the 1:1 line, whereas poorer-performing simulations exhibit larger scatter and systematic departures from this relationship.

573

574

575 A similar overall hierarchy is observed for the PI simulations. The two iLOVECLIM experiments
576 again remain among the best-performing model simulations, exhibiting low RMSE values (0.82
577 kg m⁻³), high NSE values (0.81), and KGE values of 0.86. However, MRI-CGCM3, GISS-E2-R,
578 MPI-ESM and NCAR-CCSM4 achieve comparable **performance** depending on the metric
579 considered. In particular, GISS-E2-R reaches the highest KGE (0.90), while MRI-CGCM3 shows
580 the highest NSE (0.82), illustrating once again that no single simulation systematically
581 outperforms all others across every evaluation metric.

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

585 Compared with the LGM, most simulations exhibit improved scores during the PI, particularly
586 in terms of NSE and KGE, suggesting that reproducing the pre-industrial mean state, the
587 amplitude of spatial variability, and the spatial pattern of surface density is generally less
588 challenging than under glacial boundary conditions. Nevertheless, the relative hierarchy
589 among the best- and worst-performing simulations remains broadly consistent between the
590 two periods.

591 The bias metric further highlights systematic differences among simulations. Most model
592 simulations exhibit negative biases during both periods, indicating a general tendency to
593 underestimate reconstructed surface densities. This behaviour is particularly pronounced for
594 the HadCM3 simulations during both periods and for CNRM-CM5 during the LGM. MRI-
595 CGCM3 is the only simulation displaying a slight positive bias during the PI (+0.15 kg m⁻³).
596 These systematic offsets indicate that several simulations reproduce the large-scale spatial
597 structure of surface density reasonably well while still underestimating its absolute values.

598 None of the simulations can be considered wholly inadequate, as all exhibit KGE values of at
599 least ≈ 0.5 , reflecting a moderate to good overall representation of the reconstructed surface
600 densities. The global evaluation nevertheless masks important regional contrasts. We
601 therefore examine in the following section whether the agreement identified here is
602 maintained across the different ocean basins.

603

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

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

608 We created 8 ocean basins masks (see Fig. A1). To evaluate model performance within
609 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 and its importance for major hydroclimate processes such as the monsoon and Walker circulation dynamics. Previous studies have already conducted data–model comparisons in this region, revealing its complexity.

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 is evaluated using complementary skill metrics (Bias, RMSE, NSE and KGE), providing a more direct assessment of model–data agreement.

4.1. Evaluation of model simulations performance in the North and South Indian Ocean

The Indian Ocean is divided into two dynamically distinct regions separated at the Equator (0°). The southern basin extends to 40°S, approximately corresponding to the present-day subtropical convergence and the southern limit of the Agulhas current system. This subdivision isolates the monsoon-dominated North Indian Ocean from the subtropical South Indian Ocean, allowing us to evaluate model performance in two contrasting dynamical regimes. As for the global data-model comparison, model simulation skill is assessed using four complementary metrics computed from reconstructed and simulated surface density fields: Root Mean Square Error (RMSE), Bias, Nash–Sutcliffe Efficiency (NSE) and Kling–Gupta Efficiency (KGE).

During the LGM (Fig. 10), weighted KGE values indicate that almost all simulations provide a satisfactory representation of reconstructed Indian Ocean surface density. Although some differences exist among model simulations, they remain relatively limited, suggesting broadly comparable overall performance across most simulations. In contrast, the two iLOVECLIM simulations clearly stand apart, exhibiting markedly lower KGE values and substantially poorer agreement with the reconstructions. This contrasts with their strong performance in the global LGM and PI evaluations (Fig. 9), where the two iLOVECLIM simulations ranked among the best-performing models. This discrepancy highlights the regional dependence of model–data agreement and suggests that good global performance can mask substantial regional differences.

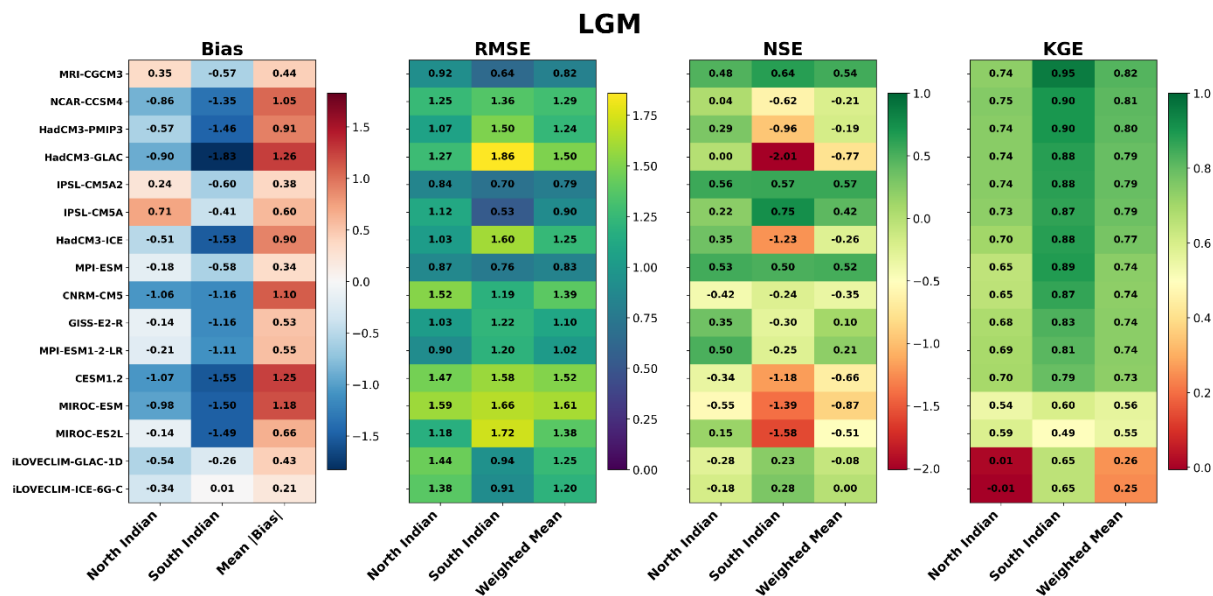


Figure 10: Simulations performance for seawater density in the North Indian and South Indian oceans during the Last Glacial Maximum (LGM). Heatmaps show four complementary performance metrics calculated from model–data comparisons at proxy reconstruction sites. Models are ranked according to their weighted 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, with values close to 0 being optimal, whereas higher values of NSE and KGE indicate better model performance, with an optimum value of 1. Negative NSE values indicate poor model performance.

This overall picture remains consistent when considering the two basins separately. KGE values are generally higher in the South Indian Ocean than in the North Indian Ocean, indicating that reconstructed density patterns are more readily reproduced in the southern Indian Ocean. Once again, the two iLOVECLIM simulations are ranked lower compare than the other simulations, particularly in the North Indian Ocean, where their KGE values approach zero. The particularly poor performance of iLOVECLIM in the North Indian Ocean likely contributes to the contrast between its global and regional skill. Because the global evaluation combines information across the entire ocean domain, strong performance in other regions may compensate for poorer agreement in the Indian Ocean.

The remaining metrics provide complementary information on different aspects of model simulations performance and broadly support the KGE assessment. Simulations associated with relatively high KGE values generally also exhibit low RMSE values, indicating limited overall model–data differences, while absolute biases remain moderate although somewhat more variable among simulations. NSE highlights larger differences in the overall agreement between simulations and the reconstructed density field, owing to its sensitivity to both variability representation and large errors. In particular, several HadCM3 simulations exhibit relatively low or even negative NSE values in the South Indian Ocean despite maintaining relatively high KGE values. This suggests that, while the simulations achieve good overall agreement with the reconstructed density field, they are affected by error structures that are more strongly penalized by the NSE than by the KGE. Conversely, the iLOVECLIM simulations display only moderate RMSE values but consistently low NSE and KGE values, highlighting that reproducing the overall magnitude of model–data differences alone is insufficient to ensure satisfactory overall model performance.

During the PI (Fig. 11), the results remain very similar to those obtained for the LGM. Most simulations again exhibit relatively high weighted KGE values, indicating satisfactory agreement with the reconstructed density field. Differences among simulations remain modest, whereas the two iLOVECLIM simulations once more stand apart, exhibiting negative KGE values that indicate a systematic inability to reproduce the reconstructed density field.

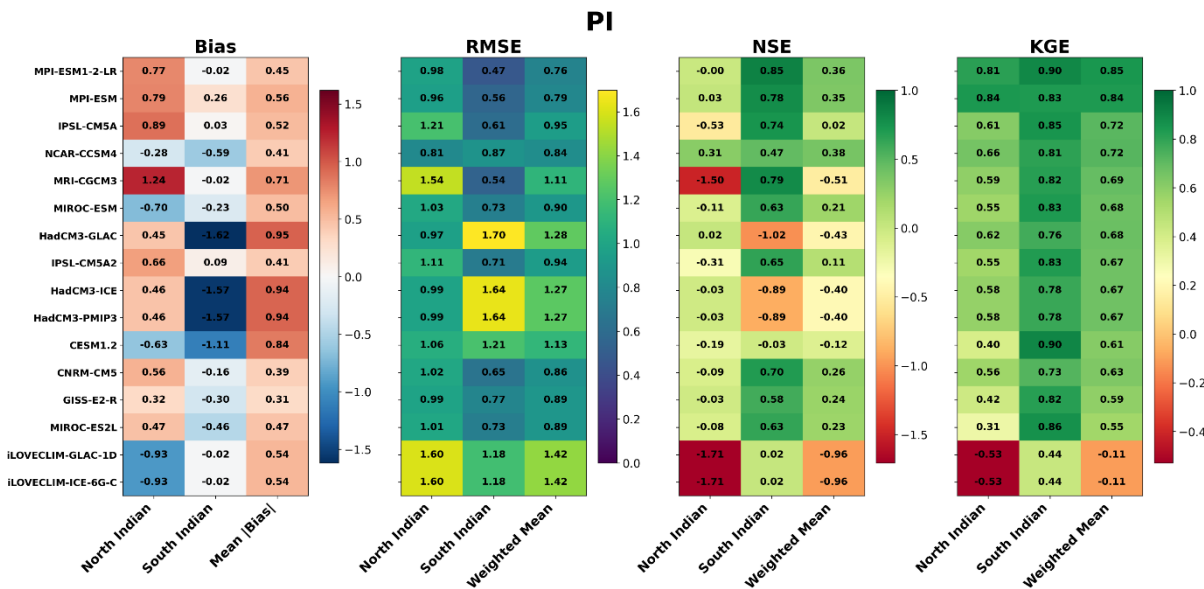


Figure 11: Simulations performance for seawater density in the North Indian and South Indian oceans during the pre-industrial period (PI). Heatmaps show four complementary performance metrics calculated from model–data comparisons at proxy reconstruction sites. Models are ranked according to their weighted mean Kling–Gupta Efficiency (KGE), averaged across both basins. The metrics shown are: mean absolute bias ($|Bias|$), 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, with values close to 0 being optimal, whereas higher values of NSE and KGE indicate better model performance, with an optimum value of 1. Negative NSE values indicate poor model performance.

The north–south contrast observed during the LGM persists during the PI. Most simulations achieve higher KGE values in the South Indian Ocean than in the North Indian Ocean, confirming that reconstructed density fields are more readily reproduced in the south. As during the LGM, the two iLOVECLIM simulations are ranked lower compare than the other simulations, with negative KGE values in the North Indian Ocean.

RMSE, Bias and NSE support the same overall interpretation while highlighting complementary aspects of model simulations performance. Overall, the four complementary metrics consistently show that most simulations reproduce reconstructed Indian Ocean surface density reasonably well during the PI, whereas the two iLOVECLIM simulations exhibit consistently poor performance across all metrics.

Beyond the North–South comparison presented in Sect. 4.1, one of the dominant modes of variability in the tropical Indian Ocean is characterized by a zonal rather than meridional structure and is most strongly expressed at interannual timescales. The Indian Ocean region, including part of the Indo-Pacific Warm Pool (IPWP), plays a critical role in the global climate system due to strong coupling between ocean and atmosphere, particularly through the Indo-Pacific Walker circulation. This circulation influences the zonal distribution of SST and thermocline depth, providing the background conditions for phenomena such as the Indian Ocean Dipole (IOD) (Saji et al., 1999; Abram et al., 2020). The Indian Ocean is also closely linked to the monsoon hydrological cycle, with precipitation patterns affecting surface salinity and, consequently, density, making it particularly sensitive to both temperature and salinity variations. However, climate models often misrepresent the mean state of the IOD, potentially leading to biases in simulating climate variability (Weller and Cai, 2013; Cai and Cowan, 2013). Notably, Abram et al. (2020) report that many models produce an overly strong thermocline–SST feedback due to a misrepresentation of the mean state, particularly an exaggerated zonal thermocline slope, which artificially increases the strength and frequency of simulated IOD events. Although research on the IOD has expanded over the past two decades, uncertainties remain regarding the controls and long-term evolution of the Indian Ocean’s mean state, especially under different climate boundary conditions. The limited timeframe of the instrumental record and persistent model biases make paleoclimate reconstructions essential for investigating past mean states and for testing model performance beyond the range of modern variability (Abram et al., 2020). To date, paleoclimate data of SST suggest that past periods, such as during the LGM, mid-Holocene or 17th century, tend to have a mean state that is more typical of a positive IOD-like, which is systematically associated with elevated IOD variability. This indicates a tight coupling between the mean state and interannual dynamics (Abram et al., 2020).

Following the basin-scale evaluation presented in Sect. 4.1, we now investigate whether the observed model–data discrepancies originate from an incorrect simulation of the tropical West–East density gradient. To do so, model performance is evaluated using a West–East density gradient specifically designed to capture the mean-state structure associated with IOD-like conditions. SST and precipitation driving SSS changes create a West–East surface salinity gradient (Fig. 12) and therefore a density gradient. Recent observations show that, at interannual timescales, SSS variability is strongly linked to the IOD and ENSO, and can interact with SST anomalies through a SST–precipitation–SSS feedback. This suggests that SSS may at times amplify rather than offset SST-related changes (Zhang et al., 2016).

In this section, the 'iLOVECLIM-ICE-6G-C' and 'iLOVECLIM-GLAC-1D' simulations have been excluded due to a salinity bias in the northern Indian Ocean region caused by excess precipitation that reduces salinity in the iLOVECLIM model in comparison to reconstructions as shown by Roche and Caley (2013) and as confirmed by skill metrics for the Indian Ocean on Figure 10 and 11.

Two boxes were defined: "Indian West", which contains 6 density reconstructions and "Indian East", which contains 5. These boxes were chosen based on the definition of the IOD by Saji et al. (1999), and slightly adjusted to ensure sufficient points within each box. Unlike the

regional North–South boxes used elsewhere in this study, these boxes intentionally straddle the Equator because they are designed to capture the tropical zonal structure of the Indian Ocean Dipole rather than hemispheric contrasts. The "Indian West" box is defined as 50°E–70°E, 12°S–12°N, corresponding to the tropical western Indian Ocean. The "Indian East" box is defined as 90°E–110°E, 10°S–2°N, corresponding to the tropical south-eastern Indian Ocean. For model–data comparisons, model values are extracted at the exact grid cells where reconstructions are available. The SST points used by Tierney et al. (2020) in the "Indian East" box are geographically close to the density reconstructions, which favours a tighter spatial match between SST and density data. In contrast, reconstructions from the MARGO (MARGO project, 2009) database in the "Indian West" box are located farther offshore, showing less spatial overlap with the available density reconstructions.

Before analyzing the LGM–piControl difference, we verified that models reproduce the West–East gradient observed during the pre-industrial period. To do this we compared the West–East gradient of SST, salinity, and surface density from each piControl simulation with EN4 observational data (1900–1999) extracted at the same grid cells as LGM reconstructions (Fig. E1). We find that all model simulations agree with the observations in terms of the gradient's sign: a negative West–East gradient for temperature, and positive gradients for salinity and density.

Temperature shows greater inter-model spread, while for salinity and density, the model median almost matches the gradient value from EN4 observations (Fig. E1). The proxy-based reconstructions West–East density gradient closely matches the EN4 observations, confirming consistency between modern observations and Late Holocene reconstructions. Regarding SST, we find a difference of around 1°C between the EN4 SST and the Tierney et al. (2020) SST reconstructions in the same locations (Fig. E1).

We examine the surface density difference between the LGM and PI periods across the two boxes (Fig. 12 (a)). The proxy-based reconstruction shows a negative West–East density anomaly, with a value close to -1 kg/m³ (Fig. 12 (a)). About half of the model simulations reproduce the correct sign, with CNRM-CM5, MRI-CGCM3, GISS-E2-R, HadCM3-GLAC1D, HadCM3-ICE-6G_C, HadCM3-PMIP3 and CESM1.2 falling within the 68% uncertainty range of the reconstruction. HadCM3-ICE-6G_C and CESM1.2 best capture both magnitude and sign.

These model–data mismatches are consistent with known CMIP-class biases: many models fail to reproduce the LGM west–east density anomaly due to errors in Walker circulation, monsoon dynamics, and regional precipitation (Feng et al., 2023; McKenna et al., 2024). Such atmospheric circulation biases propagate into the ocean (thermocline slope, zonal SST gradients) and thereby affect both temperature and salinity-driven contributions to surface density. In this light, the spread of model outcomes at the LGM can be partly attributed to differing model responses to LGM boundary conditions and to systematic CMIP biases in tropical atmospheric circulation.

To investigate whether temperature and/or salinity biases could be responsible for the model–data mismatch, we first examine the west–east gradient anomaly in temperature. On the temperature side (Fig. 12 (b) and (c)), the two datasets show differing patterns: the MARGO (MARGO project, 2009) database reports a small West–East gradient (-0.28), whereas the Tierney et al. (2020) dataset shows a positive anomaly (1.33). The SST reconstructions in MARGO (MARGO project, 2009) in this zone are mainly based on foraminiferal assemblages, whereas the reconstructions in the Tierney et al. (2020) database are mainly based on Mg/Ca proxies. As the Tierney reconstructions are geographically closer to the density records, we use them (Fig. 12 (b)) for interpreting the SST anomaly. The positive West–East gradient in Tierney et al. (2020) does not imply the western Indian Ocean was warmer than the east during the LGM; rather, it indicates a relative reduction in the zonal SST gradient compared to preindustrial conditions. This is consistent with the observed west–east density anomaly, where relatively weaker cooling in the west would contribute to lower density than in the east. SST anomalies exhibit smaller inter-model spread than density anomalies. The simulations that most closely reproduce the observed SST anomaly pattern between LGM and piControl are HadCM3-GLAC-1D, HadCM3-ICE-6G_C, HadCM3-PMIP3, and CESM1.2.

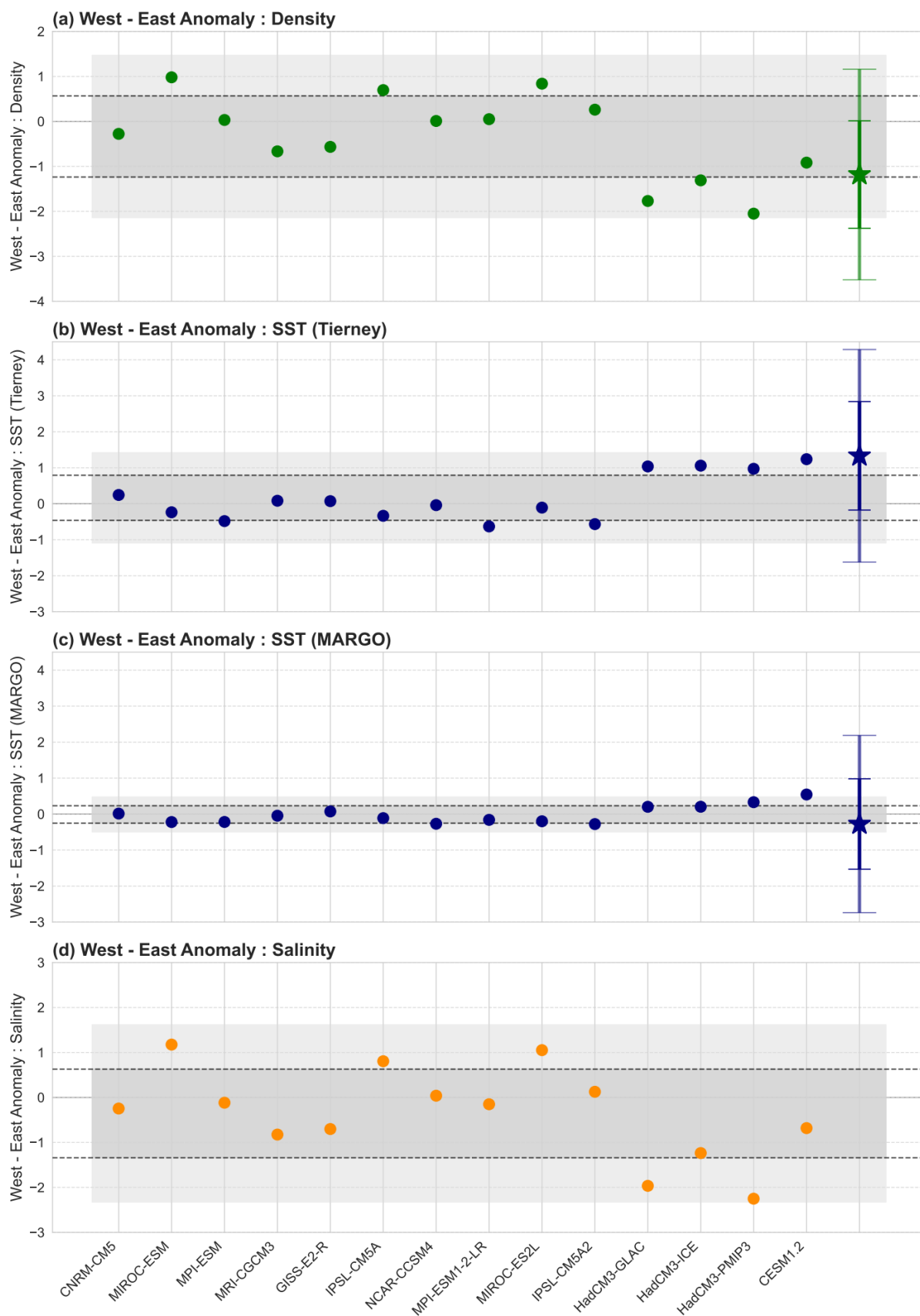


Figure 12: Model–data comparison of West–East anomalies (LGM – piControl) in the Indian Ocean. Grey shaded areas indicate model uncertainty ($\pm 1\sigma$, light shading $\pm 2\sigma$), based on simulation ensembles. Stars denote reconstruction-based anomalies and are shown for comparison only. All model values are extracted at the same grid cells as the observational sites to ensure a spatially consistent comparison. Regional uncertainties are obtained by combining uncertainties from all sediment cores from each box (West and East). For each core, the uncertainty is represented by a Gaussian standard deviation derived from its reported confidence interval. We construct inverse-variance reliability weights using two components: the within-core variance (from the CI-based σ) and the between-core spatial variance (the variance of core-specific means). The regional mean is the weighted average of the core means, and its uncertainty follows the law of total variance, ensuring that both local reconstruction uncertainty and spatial variability are preserved. To obtain smooth uncertainty bands, we then generate Monte-Carlo draws from this mixture. Uncertainty on the West–East anomaly is then obtained by standard error propagation, assuming independent regional means (square root of the sum of squared standard deviations). **(a)** Surface density anomaly (kg/m^3). Density reconstructions are derived from $\delta^{18}\text{O}_{\text{c}}$ measurements of planktonic foraminifera and converted into density estimates using the Bayesian calibration method of Caley et al. (2026). Observational uncertainty is shown by the 68% (dark green) and 95% (light green) confidence intervals. **(b)** Sea surface temperature (SST) anomaly ($^{\circ}\text{C}$) based on Tierney et al. (2020). Observational uncertainties are shown as 68% (dark blue) and 95% (blue). **(c)** Same as (b), but using the MARGO database (MARGO Project, 2009). **(d)** Sea surface salinity (SSS) anomaly (g/kg).

DiNezio et al. (2018) investigated the LGM–piControl SST anomaly in this part of the Indian Ocean using the CESM1 model, revealing a West–East SST gradient in this region. Their study showed that two main mechanisms drive the observed glacial-interglacial climate changes: first, the exposure of the Sahul shelf enhances ocean-atmosphere feedbacks that alter rainfall and temperature gradients across the Indian Ocean; second, Northern Hemisphere cooling weakens monsoonal systems by reducing moisture supply, especially over the Arabian Sea. This sensitivity is also dependent on how the newly exposed continental shelf is represented in the model. Factors such as the prescribed surface roughness, vegetation type, or albedo in the now-exposed Indonesian region can significantly impact the simulated atmospheric circulation (DiNezio and Tierney, 2013; DiNezio et al., 2018). In particular, the ability of the convection scheme to respond differently to land versus ocean surfaces plays a critical role in shaping regional precipitation patterns (Chemel et al., 2014). These model design choices likely contribute to the spread of results across PMIP3 and PMIP4 models and their varying skill in reproducing observed SST and surface density gradients.

Salinity biases were also assessed using LGM–PI anomalies and West–East gradients (Fig. 12 (d)). Without direct LGM salinity reconstructions, this relies on model outputs. Inter-model spread in salinity is comparable to density and larger than SST, emphasizing the role of freshwater fluxes and hydrological processes in shaping density changes. Models with the most negative salinity anomalies (MRI-CGCM3, GISS-E2-R, HadCM3-GLAC-1D, HadCM3-ICE-6G_C, HadCM3-PMIP3 and CESM1.2) best reproduce the observed west–east density gradient. Models failing to produce sufficiently negative salinity anomalies often have biases in monsoon precipitation location and intensity, directly affecting surface salinity. DiNezio and Tierney (2013) showed that accurate West–East salinity gradients require correct precipitation patterns. This underscores the importance of correctly simulating the Indian Ocean hydrological cycle (including major rivers) and atmospheric circulation to reproduce salinity-driven density gradients and associated climate impacts.

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).

All HadCM3 simulations follow the PMIP4 protocol for implementing LGM boundary conditions. This protocol ensures that the differences between ice-sheet reconstructions are consistently applied—not only in terms of ice-sheet mask and elevation, but also land–sea distribution, bathymetry, and far-field topography. Among them, the HadCM3-ICE-6G_C simulation provides the best agreement with the observed west–east surface density gradient. This improved performance is consistent with previous results (Izumi et al., 2023), who showed that the ICE-6G_C ice-sheet configuration induces distinct atmospheric circulation responses compared to other reconstructions. These include shifts in jet structure and stationary waves, as well as differences in surface albedo forcing and sea-ice expansion. Such large-scale circulation adjustments feedback onto the Indo-Pacific climate, which likely explains the slightly more realistic simulation of the west–east density gradient in HadCM3-ICE-6G_C.

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

We also demonstrate that this zonal SST changes are associated with zonal salinity and density changes. Of the 14 model simulations, only 7 (50%) successfully reproduce the observed West–East tropical gradient both in SST and density during the LGM, when accounting for the 68% uncertainty range of the reconstructions. This highlights current climate model limitations, indicating that key processes are insufficiently represented or that boundary condition choices, especially ice-sheet reconstructions, could influence model–data agreement and underscores the value of palaeoclimate for improving our understanding of Indian Ocean mean state under varying boundary conditions.

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.

5. Conclusion:

The quantitative density reconstruction method based on $\delta^{18}\text{O}_c$, developed by Caley et al. (2026), provides a valuable dataset for evaluating climate model simulations from PMIP3 and PMIP4 in terms of LGM density changes. Instead of comparing ensemble means, our analysis evaluates each model simulation individually.

Overall and at the global scale, models tend to slightly underestimate the magnitude of density anomalies compared to proxy-based reconstructions, which exhibit a mean anomaly of $\sim 1.6 \text{ kg/m}^3$, while model simulations average around 1.1 kg/m^3 with a range from 0.6 to 1.6 kg/m^3 . Accounting for reconstruction uncertainty improves the agreement between simulations and reconstructions, particularly by reducing discrepancies in variability. However, substantial differences between simulated and reconstructed global surface density anomaly (LGM–PI) distributions remain.

Taylor diagrams and complementary performance metrics, indicate that all simulations capture key features of the reconstructed surface densities to some extent, resulting in moderate to good overall agreement for the LGM and the PI time periods. Several simulations, including MRI-CGCM3 and IPSL-CM5A2 and the two iLOVECLIM experiments, consistently perform well across multiple global metrics. Nevertheless, this apparent global agreement masks substantial regional discrepancies revealed by the regional analyse of the Indian Ocean.

Most simulations reproduce reconstructed surface density field in the Indian Ocean reasonably well, regardless of the climatic period considered, but simulations performance is nevertheless systematically lower in the North Indian Ocean than in the South Indian Ocean. In addition, while the basin-scale density changes are reasonably reproduced, except for iLOVECLIM simulations, only 7 out of 14 remaining simulations (50%) capture the weakened West–East density gradient. This misrepresentation of the mean state of the IOD-like could affect our understanding of past IOD variability.

Discrepancies between simulated and reconstructed surface ocean density are linked to biases in SST and salinity changes, and to insufficient representation of low-latitude precipitation reductions, which are critical for reproducing density anomalies in tropical regions.

These results emphasize that evaluating individual simulations rather than ensemble means is essential, as ensemble averaging can mask inter-model differences and region-specific biases. They also demonstrate the importance of complementary diagnostics across multiple spatial scales, from global distributions to basin-scale performance metrics and finally to tropical west–east gradients. Together, these provide a more complete assessment of model skill than any single metric alone. They also highlight the need to expand LGM density reconstructions, particularly in poorly sampled open-ocean regions, to provide stronger observational constraints on model simulations. In the future, integrating these datasets within statistical observational-constraint frameworks and data assimilation approaches could help identify model simulations that most accurately reproduce past climate states and could ultimately improve confidence in future climate projections.

Appendices:

A. Study basins:

We first considered the major oceanic basins: The Atlantic, Indian, and Pacific Oceans, the circumpolar Southern Ocean. To better capture regional processes and analyze these regions more precisely, some of these large basins were subdivided into northern and southern parts, resulting in a total of eight ocean basins. Specifically, the North Atlantic, following the density reconstruction method of Caley et al. (2026), extends from 0° to 40°N, while the South Atlantic is restricted from 0°N to 40°S. Similarly, the North Indian Basin extends from 0°S northward, and the South Indian and Pacific Basins are restricted from 0°N to 40°S. The North Pacific is separated from the North of 40°N basin at 40°N of latitude.

The Mediterranean Sea was excluded from the analysis, as this region is particularly difficult for climate models to simulate and the limited number of proxy data points resulted in statistically insignificant regressions.

The study basins were defined using the NOAA WOA23 mask file (1°×1° grid: https://www.ncei.noaa.gov/data/oceans/woa/WOA23/MASKS/basinmask_01.msk), with modifications applied to create these eight final regions (Fig. A1).

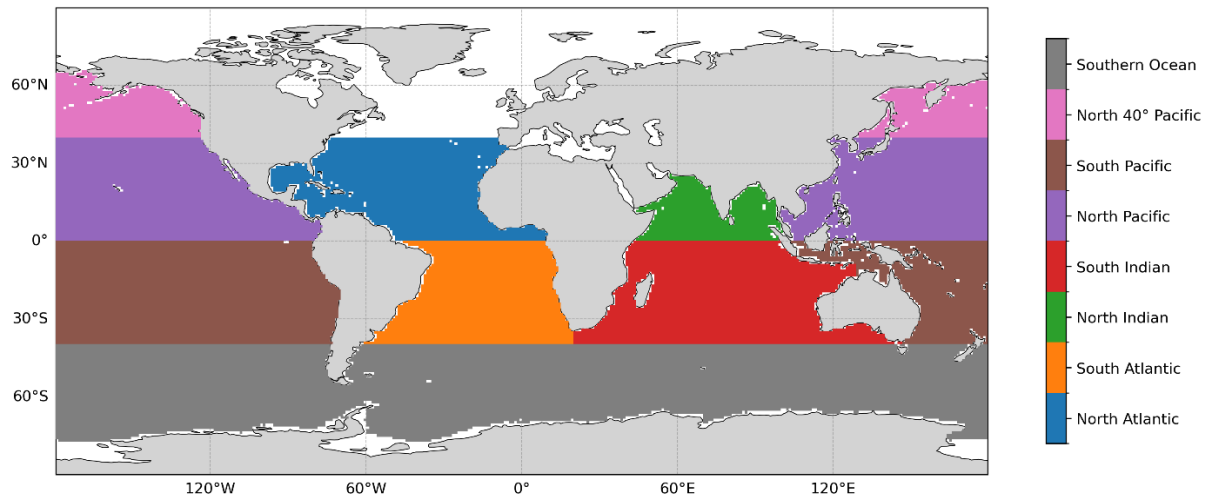


Figure A1: Map of selected basins in this study. The study basins were created from the NOAA WOA23 mask file (which has a 1°×1° grid) ([website https://www.ncei.noaa.gov/data/oceans/woa/WOA23/MASKS/basinmask_01.msk](https://www.ncei.noaa.gov/data/oceans/woa/WOA23/MASKS/basinmask_01.msk)) then some basins 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.

B. Latitudinal profiles of surface density, SSS and SST anomalies (LGM-PI):

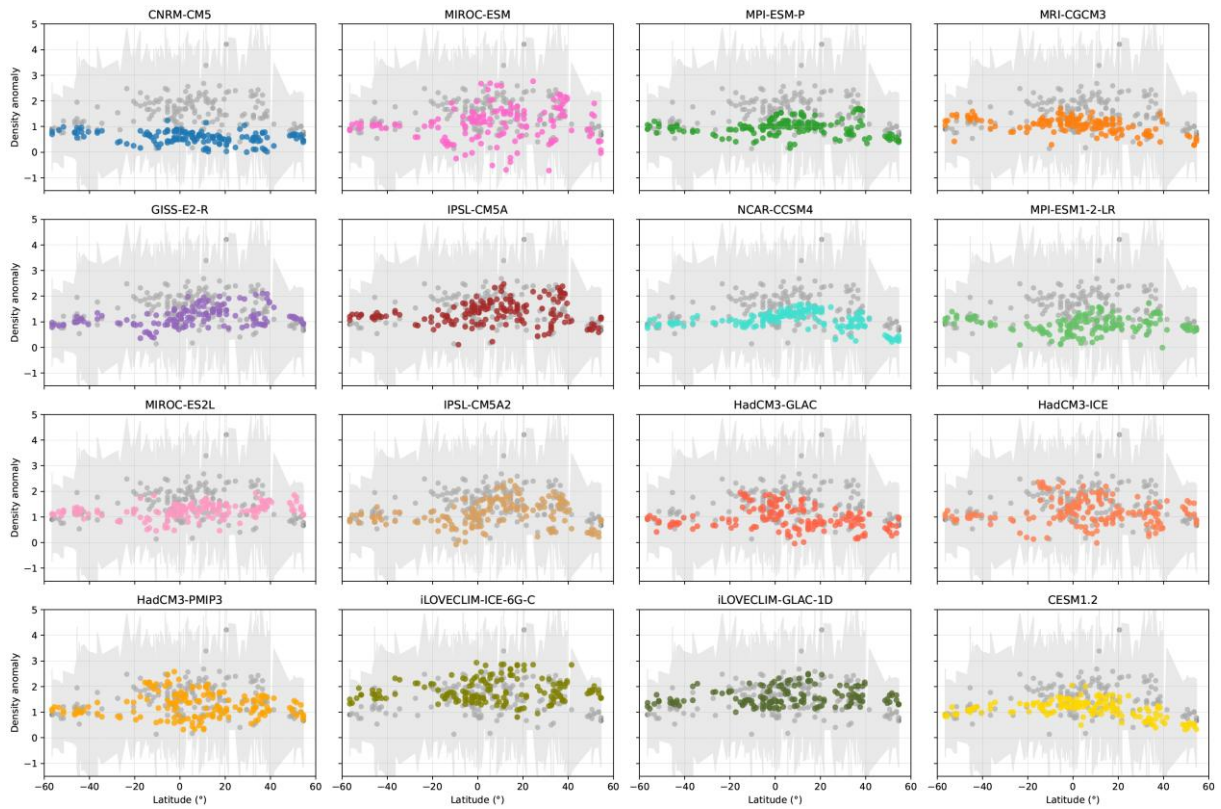


Figure B1: Density anomaly (kg/m^3) as a function of latitude for each model simulation (colored dots), compared with observational data (grey dots) and the 95% confidence interval (grey shading). Model outliers, identified using the interquartile range (IQR) method (values outside $1.5 \times \text{IQR}$), were excluded to minimize the influence of extreme values and highlight robust spatial patterns. This filtering enhances the main latitudinal structure and modelled density anomalies while preserving all available latitude sampling points.

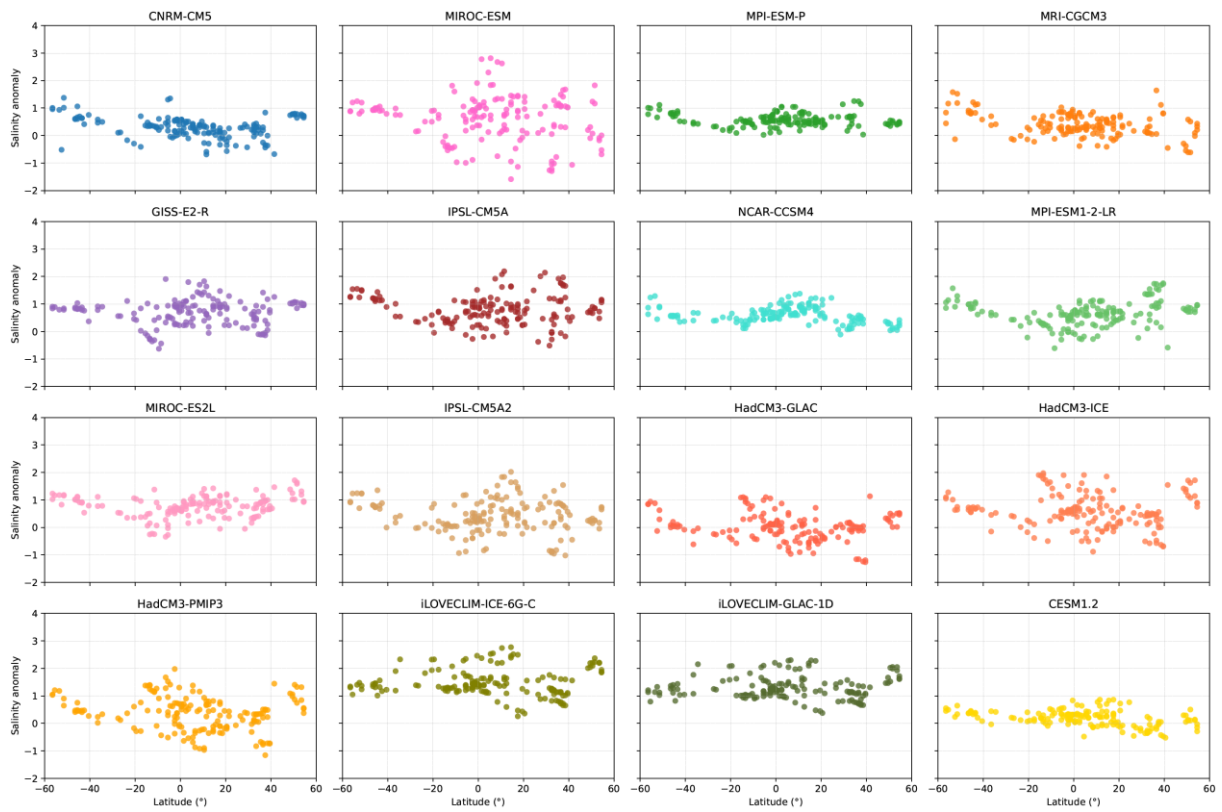
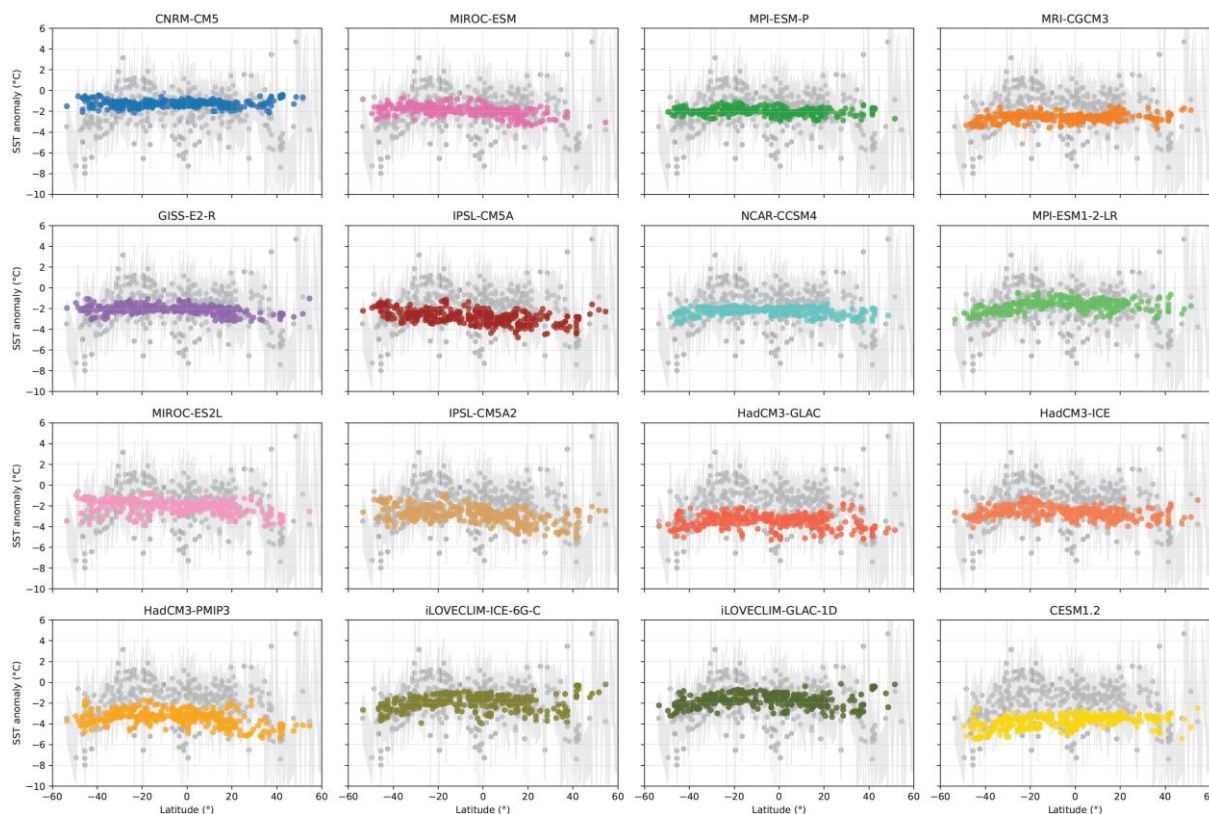


Figure B2: Zonal distribution of salinity anomalies (LGM-PI) simulated by each model (colored dots). Model outliers, identified using the interquartile range (IQR) method (values outside $1.5 \times \text{IQR}$), were excluded to minimize the influence of extreme values and highlight consistent spatial patterns. This filtering emphasizes the latitudinal structure of simulated salinity anomalies across models while preserving all available sampling points.



970 **Figure B3:** Zonal distribution of SST anomalies (LGM-PI) from the MARGO (MARGO project, 2009) reconstruction (grey dots)
 971 and $\pm 1\sigma$ uncertainties (grey shading), compared with model simulations (colored dots). Model outliers, identified using the
 972 interquartile range (IQR) method (values outside $1.5 \times \text{IQR}$), were excluded to minimize the impact of extreme values and
 973 highlight robust latitudinal patterns. This filtering enhances the readability of the large-scale SST anomaly structure while
 974 preserving all available latitude points.

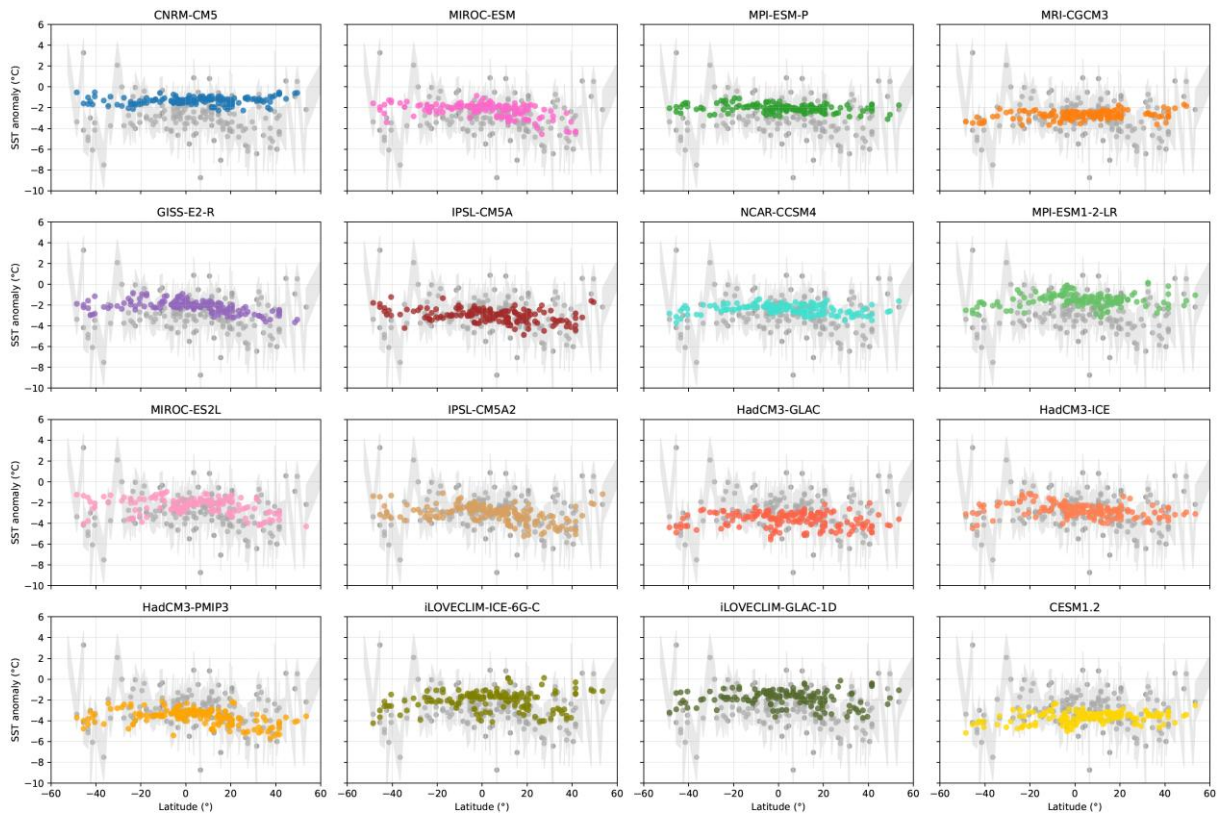


Figure B4: Zonal distribution of SST anomalies (LGM-PI) from Tierney et al. (2020) reconstruction (grey dots) and $\pm 1\sigma$ uncertainties (grey shading), compared with model simulations (colored dots). Model outliers, identified using the interquartile range (IQR) method (values outside $1.5 \times \text{IQR}$), were excluded to minimize the impact of extreme values and highlight consistent spatial patterns. This filtering clarifies the latitudinal structure of simulated SST anomalies while preserving all available observation-model comparison points.

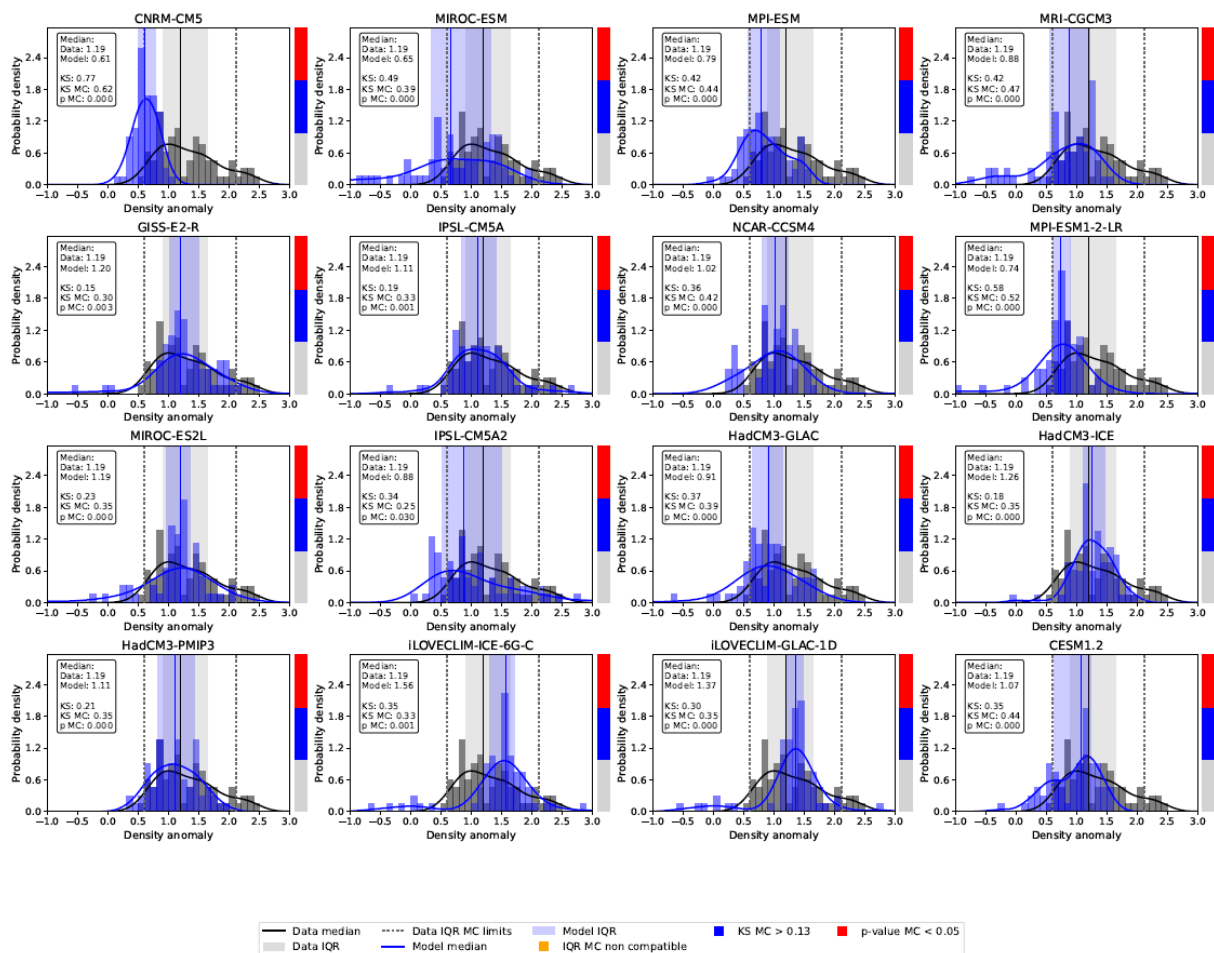


Figure C1: Distribution histograms of surface density anomalies (LGM-PI, kg/m³) at the global scale. Density reconstructions are shown in black, model simulations in blue. Kernel Density Estimates (KDEs) illustrate the central tendency and overall shape of the distributions. Vertical lines indicate the 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.

D. Global Evaluation of Surface Density: Models vs. Reconstructions (PI & LGM):

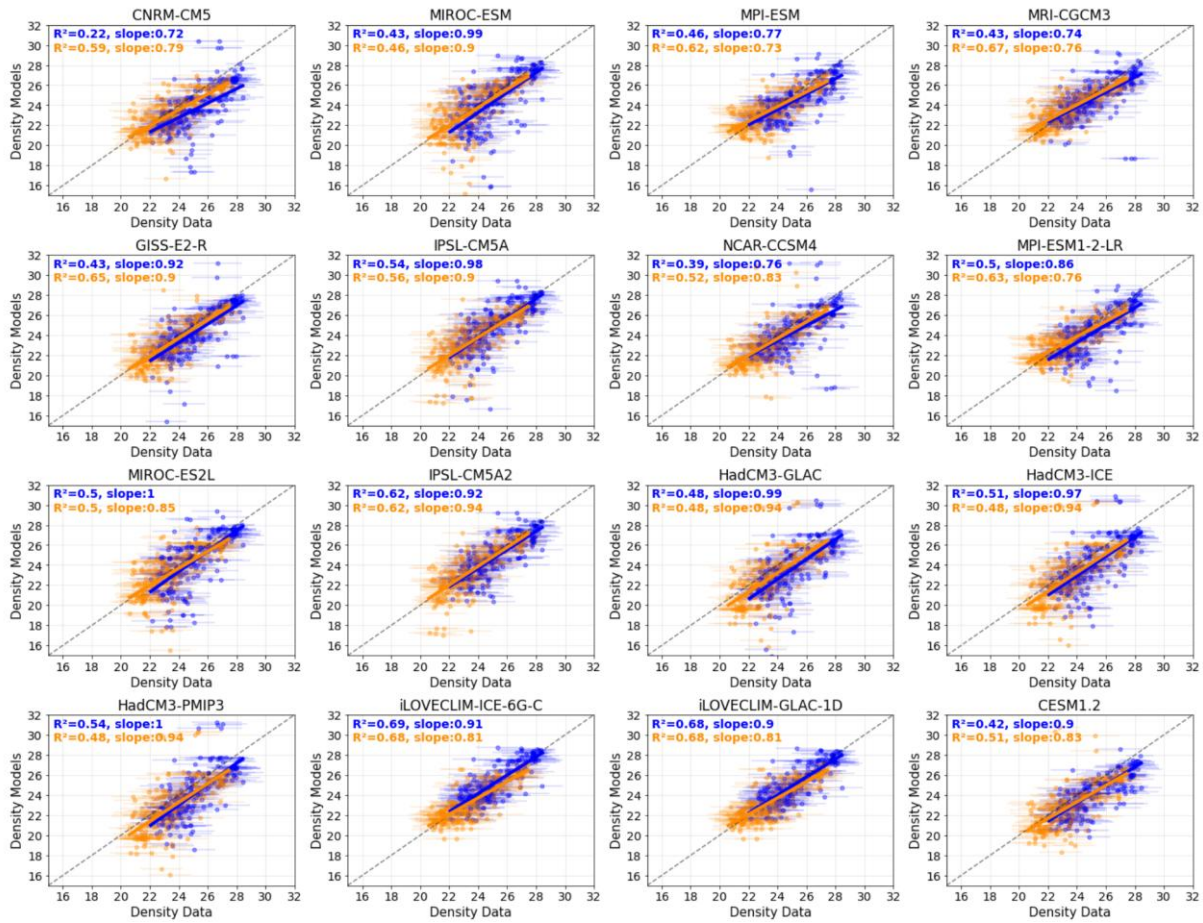


Figure D1: Linear regressions between absolute surface density (kg/m^3) from proxy-based reconstructions (x-axis) and model simulations (y-axis), aggregated at the global scale (across all selected basins). Results are shown for the LGM period (blue) and the piControl period (orange). Error bars on the x-axis represent the 95% confidence intervals of the reconstructed values. The slope and R^2 values correspond to standard linear regressions, accounting for uncertainties on the x-axis (the Monte Carlo method was applied here). All regressions shown are statistically significant ($p < 0.05$).

E. Focus on the Indian Ocean:

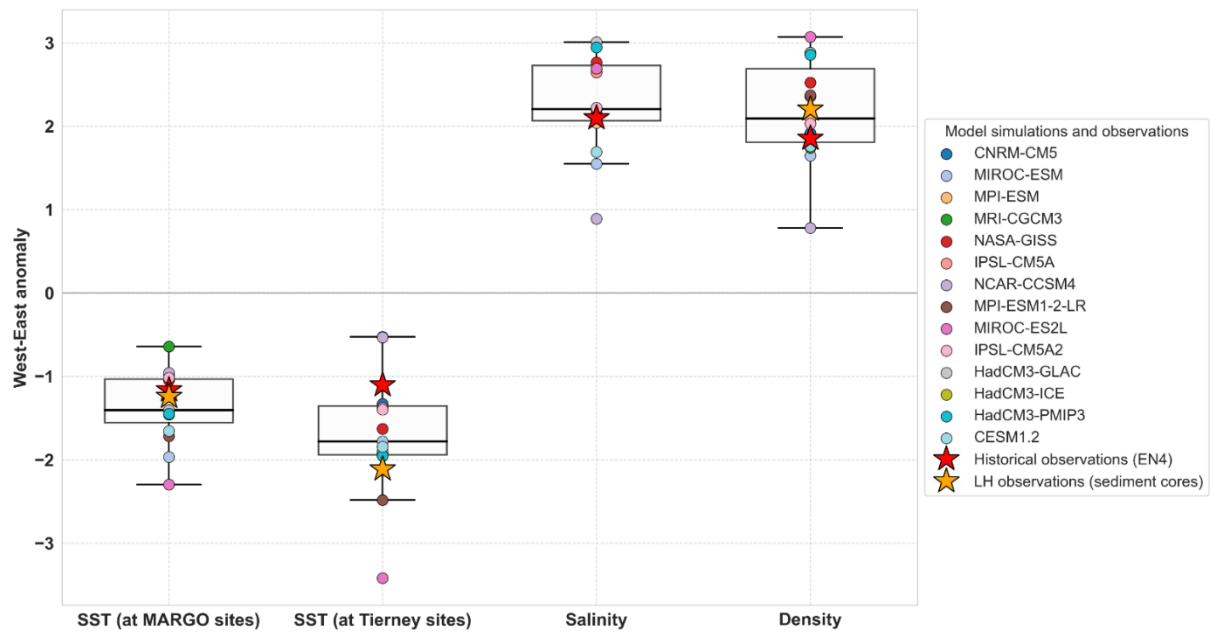


Figure E1: Boxplots showing West-East anomalies (Indian_East – Indian_West) in Sea Surface Temperature ($^{\circ}\text{C}$), (from the MARGO and Tierney datasets, separately), Sea Surface Salinity (g/kg), and sea surface density (kg/m^3) during the pre-industrial period. The boxplots represent the distribution 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)). The "Indian West" box is defined as 50°E – 70°E , 12°S – 12°N , corresponding to the tropical western Indian Ocean. The "Indian East" box is defined as 90°E – 110°E , 10°S – 2°N , corresponding to the tropical south-eastern Indian Ocean.

Data availability:

Most of the CMIP5/PMIP3 and CMIP6/PMIP4 climate model simulations used in this study are publicly available through the Earth System Grid Federation (ESGF). The simulations IPSL-CM5A2, HadCM3-GLAC-1D, HadCM3-ICE-6G_C, HadCM3-PMIP3, iLOVECLIM-GLAC-1D, iLOVECLIM-ICE-6G_C, and CESM1.2 are not available via ESGF. CESM1.2 simulation outputs are openly available from Zenodo (<https://zenodo.org/records/14957995>). The HadCM3 model simulations can be accessed at <https://www.paleo.bristol.ac.uk/ummodel/scripts/papers/>. iLOVECLIM simulations outputs are available upon request from Nathalie Bouttes, and IPSL-CM5A2 data can be obtained by contacting Masa Kageyama. References for all simulations are provided in Table 1.

The $\delta^{18}\text{O}_c$ database and the Python code to compute surface ocean density are related to Caley et al. 2026 and can be found at https://github.com/nicrie/density_uncertainty/tree/main/data

Author contribution:

TC and HB designed the study. HB and TC designed the analyses, and HB performed them. MK, PB and DS assisted in retrieving model simulations data. HB and TC analysed the results with contribution and discussion of all co-authors. HB produced the figures and wrote the article with help from TC and input from all co-authors.

Competing interests:

The authors declare that they have no competing interest.

Acknowledgements:

We thank Valentin Portmann for his support and expertise in statistics and Python-based data processing, and for providing the regridded EN4 dataset used in this study. We also thank Didier Roche for carefully reviewing the manuscript and providing constructive feedback. We are grateful to the paleoclimate modelling community for providing access to model outputs used in this study, and we warmly thank Masa Kageyama, Jiang Zhu, Nathaëlle Bouttes, and Ruza Ivanovic for their help in retrieving PMIP simulations.

The authors would also like to thank all the modelling groups who provided the PMIP3 and PMIP4 outputs used in this analysis, as well as the CMIP panel and ESGF infrastructures for making the data available, and WCRP and CLIVAR for supporting the PMIP project. Héloïse Barathieu acknowledges the use of the IPSL (ESPRI – Ensemble de Services Pour la Recherche l’IPSL – computing and data centre (<https://mesocentre.ipsl.fr/>, last access: July 2025).

Financial support:

This research was supported by the ANR HYDRATE project (grant no. ANR-21-CE01-0001) of the French Agence Nationale de la Recherche.

References:

- Abram, N. J., Hargreaves, J. A., Wright, N. M., Thirumalai, K., Ummenhofer, C. C., and England, M. H.: Palaeoclimate perspectives on the Indian Ocean dipole, *Quat. Sci. Rev.*, 237, 106302, <https://doi.org/10.1016/j.quascirev.2020.106302>, 2020.
- Adloff, M., Reick, C. H., and Claussen, M.: Earth system model simulations show different feedback strengths of the terrestrial carbon cycle under glacial and interglacial conditions, *Earth Syst. Dynam.*, 9, 413–425, <https://doi.org/10.5194/esd-9-413-2018>, 2018.
- Annan, J. D., Hargreaves, J. C., and Mauritsen, T.: A new global surface temperature reconstruction for the Last Glacial Maximum, *Clim. Past*, 18, 1883–1898, <https://doi.org/10.5194/cp-18-1883-2022>, 2022.
- Ayache, M., Swingedouw, D., Mary, Y., Eynaud, F., and Colin, C.: Multi-centennial variability of the AMOC over the Holocene: A new reconstruction based on multiple proxy-derived SST records, *Glob. Planet. Change*, 170, 172–189, <https://doi.org/10.1016/j.gloplacha.2018.08.016>, 2018.
- Bouttes, N., Lhardy, F., Quiquet, A., Paillard, D., Goosse, H., and Roche, D. M.: Deglacial climate changes as forced by different ice sheet reconstructions, *Clim. Past*, 19, 1027–1042, <https://doi.org/10.5194/cp-19-1027-2023>, 2023.
- Braconnot, P., Harrison, S. P., Kageyama, M., Bartlein, P. J., Masson-Delmotte, V., Abe-Ouchi, A., Otto-Bliesner, B., and Zhao, Y.: Evaluation of climate models using palaeoclimatic data, *Nat. Clim. Change*, 2, 417–424, <https://doi.org/10.1038/nclimate1456>, 2012.
- Brady, E. C., Otto-Bliesner, B. L., Kay, J. E., and Rosenbloom, N.: Sensitivity to glacial forcing in the CCSM4, *J. Clim.*, 26, 1901–1925, <https://doi.org/10.1175/JCLI-D-11-00416.1>, 2013.
- Brierley, C., Thirumalai, K., Grindrod, E., and Barnsley, J.: Indian Ocean variability changes in the Paleoclimate Modelling Intercomparison Project, *Clim. Past*, 19, 681–701, <https://doi.org/10.5194/cp-19-681-2023>, 2023.
- Cai, W., and Cowan, T.: Why is the amplitude of the Indian Ocean Dipole overly large in CMIP3 and CMIP5 climate models?, *Geophys. Res. Lett.*, 40, 1200–1205, <https://doi.org/10.1002/grl.50208>, 2013.
- 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.
- Chemel, C., Russo, M. R., Hosking, J. S., Telford, P. J., and Pyle, J. A.: Sensitivity of tropical deep convection in global models: effects of horizontal resolution, surface constraints, and 3D atmospheric nudging, *Atmos. Sci. Lett.*, 16, 148–154, <https://doi.org/10.1002/asl2.540>, 2014.
- 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., Fichet, T., Flavoni, S., Guez, L., Guilyardi, E., Hourdin, F., Ghattas, J., Kageyama, M., Khodri, M., Labetoulle, S.,

1106 Lefebvre, M.-P., Levy, C., Li, L., Lott, F., Madec, G., Marchand, M., Meurdesoif, Y., Rio, C.,
 1107 Schulz, M., Swingedouw, D., Szopa, S., Viovy, N., and Vuichard, N.: IPSL-CM5A-LR model output
 1108 prepared for CMIP5 piControl experiment, served by ESGF, World Data Center for Climate
 1109 (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.IPILpc, 2016.

1110 DiNezio, P. N., and Tierney, J. E.: The effect of sea level on glacial Indo-Pacific climate, *Nat.*
 1111 *Geosci.*, 6, 485–491, <https://doi.org/10.1038/ngeo1823>, 2013.

1112 DiNezio, P. N., Tierney, J. E., Otto-Bliesner, B. L., Timmermann, A., Bhattacharya, T.,
 1113 Rosenbloom, N., and Brady, E.: Glacial changes in tropical climate amplified by the Indian
 1114 Ocean, *Sci. Adv.*, 4, eaat9658, <https://doi.org/10.1126/sciadv.aat9658>, 2018.

1115 Douville, H., Raghavan, K., Renwick, J., Allan, R. P., Arias, P. A., Barlow, M., Cerezo-Mota, R.,
 1116 Cherchi, A., Gan, T. Y., Gergis, J., Jiang, D., Khan, A., Pokam Mba, W., Rosenfeld, D., Tierney, J.,
 1117 and Zolina, O.: Water Cycle Changes, in: *Climate Change 2021 – The Physical Science Basis*,
 1118 Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental
 1119 Panel on Climate Change, edited by: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L.,
 1120 Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K.,
 1121 Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B.,
 1122 Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1055–1210,
 1123 <https://doi.org/10.1017/9781009157896.010>, 2023.

1124 Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.:
 1125 Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design
 1126 and organization, *Geosci. Model Dev.*, 9, 1937–1958, [https://doi.org/10.5194/gmd-9-1937-](https://doi.org/10.5194/gmd-9-1937-2016)
 1127 2016, 2016.

1128 Eyring, V., Gillett, N. P., Achuta Rao, K. M., Barimalala, R., Barreiro Parrillo, M., Bellouin, N.,
 1129 Cassou, C., Durack, P. J., Kosaka, Y., McGregor, S., Min, S., Morgenstern, O., and Sun, Y.: Human
 1130 Influence on the Climate System, in: *Climate Change 2021: The Physical Science Basis*.
 1131 Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental
 1132 Panel on Climate Change, edited by: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L.,
 1133 Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K.,
 1134 Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B.,
 1135 Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 423–552,
 1136 <https://doi.org/10.1017/9781009157896.005>, 2021.

1137 Feng, J., Lian, T., and Chen, D.: Tropical Indian Ocean mixed layer bias in CMIP6 CGCMs
 1138 primarily attributed to the AGCM surface wind bias, *J. Climate*, 36, 4169–4188,
 1139 <https://doi.org/10.1175/JCLI-D-22-0546.1>, 2023.

1140 Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., Cox, P., Driouech,
 1141 F., Emori, S., Eyring, V., Forest, C., Gleckler, P., Guilyardi, E., Jakob, C., Kattsov, V., Reason, C.,
 1142 and Rummukainen, M.: Evaluation of climate models, in: *Climate Change 2013: The Physical*
 1143 *Science Basis*. Contribution of Working Group I to the Fifth Assessment Report of the
 1144 Intergovernmental Panel on Climate Change, edited by: Stocker, T. F., Qin, D., Plattner, G.-K.,
 1145 Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P. M., Cambridge

University Press, Cambridge, United Kingdom and New York, NY, USA, 741–866,
<https://doi.org/10.1017/CBO9781107415324.020>, 2014.

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D., Neale, R. B., Rasch, P. J., Vertenstein, M., Worley, P., Yang, Z., and Zhang, M.: The community climate system model version 4, *J. Climate*, 24, 4973–4991, <https://doi.org/10.1175/2011JCLI4083.1>, 2011.

Good, S. A., Martin, M. J., and Rayner, N. A.: EN4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates, *J. Geophys. Res. Oceans*, 118, 6704–6716, <https://doi-org.docelec.u-bordeaux.fr/10.1002/2013JC009067>, 2013.

Hajima, T., Watanabe, M., Yamamoto, A., Tatebe, H., Noguchi, M. A., Abe, M., Kato, M., Kawamiya, M., and Takemura, T.: Development of the MIROC-ES2L Earth system model and the evaluation of biogeochemical processes and feedbacks, *Geosci. Model Dev.*, 13, 2197–2244, <https://doi.org/10.5194/gmd-13-2197-2020>, 2020.

Hajima, T., Abe, M., Arakawa, O., Suzuki, T., Komuro, Y., Ogura, T., Ogochi, 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.

Harrison, S. P., Bartlein, P. J., Brewer, S., Prentice, I. C., Boyd, M., Hessler, I., Otto-Bliesner, B., Brady, E., Foley, K., and Willis, K.: Climate model benchmarking with glacial and mid-Holocene climates, *Clim. Dyn.*, 43, 671–688, <https://doi.org/10.1007/s00382-013-1922-6>, 2014.

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.

Izumi, K., Valdes, P., Ivanovic, R., and Gregoire, L.: Impacts of the PMIP4 ice sheets on Northern Hemisphere climate during the last glacial period, *Clim. Dyn.*, 60, 2481–2499, <https://doi.org/10.1007/s00382-022-06456-1>, 2023.

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.

Joussaume, S., and Taylor, K. E.: Status of the Paleoclimate Modeling Intercomparison Project (PMIP), in: *Proceedings of the first international AMIP scientific conference*, WRCP Report, 425–430, 1995.

1186 Jungclaus, J., Giorgetta, M., Reick, C., Legutke, S., Brovkin, V., Crueger, T., Esch, M., Fieg, K.,
 1187 Fischer, N., Glushak, K., Gayler, V., Haak, H., Hollweg, H.-D., Kinne, S., Kornblueh, L., Matei, D.,
 1188 Mauritsen, T., Mikolajewicz, U., Müller, W., Notz, D., Pohlmann, T., Raddatz, T., Rast, S.,
 1189 Roeckner, E., Salzmann, M., Schmidt, H., Schnur, R., Segschneider, J., Six, K., Stockhause, M.,
 1190 Wegner, J., Widmann, H., Wieners, K.-H., Claussen, M., Marotzke, J., and Stevens, B.: cmip5
 1191 output2 MPI-M MPI-ESM-P piControl, served by ESGF, World Data Center for Climate (WDCC)
 1192 at DKRZ, <https://hdl.handle.net/21.14106/ce74e48c66620583cc8c1fe6b607de36a3d6ddb8>,
 1193 2014.

1194 Jungclaus, J., Mikolajewicz, U., Kapsch, M.-L., D’Agostino, R., Wieners, K.-H., Giorgetta, M.,
 1195 Reick, C., Esch, M., Bittner, M., Legutke, S., Schupfner, M., Wachsmann, F., Gayler, V., Haak,
 1196 H., de Vrese, P., Raddatz, T., Mauritsen, T., von Storch, J.-S., Behrens, J., Brovkin, V., Claussen,
 1197 M., Crueger, T., Fast, I., Fiedler, S., Hagemann, S., Hohenegger, C., Jahns, T., Kloster, S., Kinne,
 1198 S., Lasslop, G., Kornblueh, L., Marotzke, J., Matei, D., Meraner, K., Modali, K., Müller, W.,
 1199 Nabel, J., Notz, D., Peters-von Gehlen, K., Pincus, R., Pohlmann, H., Pongratz, J., Rast, S.,
 1200 Schmidt, H., Schnur, R., Schulzweida, U., Six, K., Stevens, B., Voigt, A., and Roeckner, E.: MPI-
 1201 M MPI-ESM1.2-LR model output prepared for CMIP6 PMIP lgm, Earth System Grid Federation,
 1202 doi: 10.22033/ESGF/CMIP6.6642, 2019.

1203 Kageyama, M., Denvil, S., Foujols, M.-A., Caubel, A., Marti, O., Dufresne, J.-L., Bopp, L., Cadule,
 1204 P., Ethé, C., Idelkadi, A., Mancip, M., Masson, S., Mignot, J., Musat, I., Balkanski, Y., Bekki, S.,
 1205 Bony, S., Braconnot, P., Brockman, P., Codron, F., Cozic, A., Cugnet, D., Fairhead, L., Fichefet,
 1206 T., Flavoni, S., Guez, L., Guilyardi, E., Hourdin, F., Ghattas, J., Khodri, M., Labetoulle, S.,
 1207 Lefebvre, M.-P., Levy, C., Li, L., Lott, F., Madec, G., Marchand, M., Meurdesoif, Y., Rio, C.,
 1208 Schulz, M., Swingedouw, D., Szopa, S., Viovy, N., and Vuichard, N.: IPSL-CM5A-LR model output
 1209 prepared for CMIP5 lgm experiment, served by ESGF, World Data Center for Climate (WDCC)
 1210 at DKRZ, doi: 10.1594/WDCC/CMIP5.IPILlg, 2016.

1211 Kageyama, M., Albani, S., Braconnot, P., Harrison, S. P., Hopcroft, P. O., Ivanovic, R. F.,
 1212 Joussaume, S., Lora, J. M., Mikolajewicz, U., Roche, D. M., Volodin, E., and Zheng, W.: The
 1213 PMIP4 contribution to CMIP6—Part 4: Scientific objectives and experimental design of the
 1214 PMIP4-CMIP6 Last Glacial Maximum experiments and PMIP4 sensitivity experiments, *Geosci.*
 1215 *Model Dev.*, 10, 4035–4055, <https://doi.org/10.5194/gmd-10-4035-2017>, 2017.

1216
 1217 Kageyama, M., Braconnot, P., Harrison, S. P., Haywood, A. M., Jungclaus, J. H., Otto-Bliesner,
 1218 B. L., Rachmayani, R., Zheng, W., and Zhou, T.: The PMIP4 contribution to CMIP6—Part 1:
 1219 Overview and over-arching analysis plan, *Geosci. Model Dev.*, 11, 1033–1057,
 1220 <https://doi.org/10.5194/gmd-11-1033-2018>, 2018.

1221 Kageyama, M., Harrison, S. P., Kapsch, M.-L., Lofverstrom, M., Lora, J. M., Mikolajewicz, U.,
 1222 Sherriff-Tadano, S., Vadsaria, T., Abe-Ouchi, A., Bouttes, N., Chandan, D., Gregoire, L. J.,
 1223 Ivanovic, R. F., Izumi, K., LeGrande, A. N., Lhardy, F., Lohmann, G., Morozova, P. A., Ohgaito,
 1224 R., Paul, A., Peltier, W. R., Poulsen, C. J., Quiquet, A., Roche, D. M., Shi, X., Tierney, J. E., Valdes,
 1225 P. J., Volodin, E., and Zhu, J.: The PMIP4 Last Glacial Maximum experiments: preliminary
 1226 results and comparison with the PMIP3 simulations, *Clim. Past*, 17, 1065–1089,
 1227 <https://doi.org/10.5194/cp-17-1065-2021>, 2021

1228 Kageyama, M., Braconnot, P., Chiessi, C. M., Rehfeld, K., Ait Brahimi, Y., Dütsch, M., Gwinneth,
1229 B., Hou, A., Loutre, M.-F., Hendrigan, M., Meissner, K., Mongwe, P., Otto-Bliesner, B., Pezzi, L.
1230 P., Rovere, A., Seltzer, A., Sime, L., and Zhu, J.: Lessons from paleoclimates for recent and
1231 future climate change: opportunities and insights, *Front. Clim.*, 6, 1511997,
1232 <https://doi.org/10.3389/fclim.2024.1511997>, 2024

1233

1234 Lhardy, F., Bouttes, N., Roche, D. M., Crosta, X., Waelbroeck, C., and Paillard, D.: Impact of
1235 Southern Ocean surface conditions on deep ocean circulation during the LGM: a model
1236 analysis, *Clim. Past*, 17, 1139–1159, <https://doi.org/10.5194/cp-17-1139-2021>, 2021.

1237

1238 MARGO Project Members: Constraints on the magnitude and patterns of ocean cooling at
1239 the Last Glacial Maximum, *Nat. Geosci.*, 2, 127–132, <https://doi.org/10.1038/NGEO411>, 2009.

1240

1241 Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., Crueger, T., Esch, M.,
1242 Fast, I., Fiedler, S., Hagemann, S., Hedemann, C., Hohenegger, C., Ilyina, T., Kornblueh, L.,
1243 Lohmann, K., Mäkelä, J., Meraner, K., Mikolajewicz, U., Modali, K., Müller, W. A., Nabel, J. E.
1244 M. S., Nam, C. C. W., Notz, D., Pincus, R., Pohlmann, H., Pongratz, J., Popp, M., Raddatz, T.,
1245 Rast, S., Redler, R., Reick, C. H., Rohrschneider, T., Schemann, V., Schmidt, H., Schnur, R.,
1246 Schulzweida, U., Six, K. D., Stein, L., Stemmler, I., Stevens, B., Storch, J. S., Tian, F., Voigt, A.,
1247 Vrese, P., Wieners, K.-H., Wilkenskjaeld, S., Winkler, A., and Roeckner, E.: Developments in the
1248 MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and its response to increasing CO₂, *J.*
1249 *Adv. Model. Earth Syst.*, 11, 998–1038, <https://doi.org/10.1029/2018MS001400>, 2019.

1250

1251 McKenna, S., Santoso, A., Sen Gupta, A., and Taschetto, A. S.: Understanding biases in Indian
1252 Ocean seasonal SST in CMIP6 models, *J. Geophys. Res. Oceans*, 129, e2023JC020330,
1253 <https://doi.org/10.1029/2023JC020330>, 2024.

1254

1255 Mix, A. C., Bard, E., and Schneider, R.: Environmental processes of the ice age: land, oceans,
1256 glaciers (EPILOG), *Quat. Sci. Rev.*, 20, 627–657, [https://doi.org/10.1016/S0277-3791\(00\)00145-1](https://doi.org/10.1016/S0277-3791(00)00145-1), 2001.

1257

1258 NASA/GISS: NASA-GISS: GISS-E2-R model output prepared for CMIP5 last glacial maximum,
1259 served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi:
1260 10.1594/WDCC/CMIP5.GIGRIg, 2014.

1261

1262 NASA/GISS: NASA-GISS: GISS-E2-R model output prepared for CMIP5 pre-industrial control,
1263 served by ESGF, World Data Center for Climate (WDCC) at DKRZ, doi:
1264 10.1594/WDCC/CMIP5.GIGRpc, 2014.

1265

1266 NOAA: WOA23 basin mask file (1°×1° grid). National Centers for Environmental Information,
1267 NOAA, USA.
1268 https://www.ncei.noaa.gov/data/oceans/woa/woa23/MASKS/basinmask_01.msk, accessed
1269 2025.

1267 NODC: World Ocean Atlas 1998 (WOA98). National Oceanographic Data Center, Silver Spring,
1268 MD, USA. <https://psl.noaa.gov/data/gridded/data.nodc.woa98.html>, 1998.

1269 Ohgaito, R., Abe-Ouchi, A., Abe, M., Arakawa, O., Ogochi, K., Hajima, T., Watanabe, M.,
1270 Yamamoto, A., Tatebe, H., Noguchi, M. A., Ito, A., Yamazaki, D., Ito, A., Takata, K., Watanabe,
1271 S., Kawamiya, M., and Tachiiri, K.: MIROC MIROC-ES2L model output prepared for CMIP6 PMIP
1272 lgm, Earth System Grid Federation, doi: 10.22033/ESGF/CMIP6.5644, 2019.

1273 Roche, D. M., and Caley, T.: $\delta^{18}\text{O}$ water isotope in the iLOVECLIM model (version 1.0) – Part 2:
1274 Evaluation of model results against observed $\delta^{18}\text{O}$ in water samples, *Geosci. Model Dev.*, 6,
1275 1493–1504, <https://doi.org/10.5194/gmd-6-1493-2013>, 2013.

1276 Roquet, F., Madec, G., McDougall, T. J., and Barker, P. M.: Accurate polynomial expressions
1277 for the density and specific volume of seawater using the TEOS-10 standard, *Ocean Model.*,
1278 90, 29–43, <https://doi.org/10.1016/j.ocemod.2015.04.002>, 2015.

1279 Saji, N. H., Goswami, B. N., Vinayachandran, P. N., and Yamagata, T.: A dipole mode in the
1280 tropical Indian Ocean, *Nature*, 401, 360–363, <https://doi.org/10.1038/43854>, 1999.

1281 Schmidt, G. A., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., Bauer, M., Bauer,
1282 S. E., Bhat, M. K., Bleck, R., Canuto, V., Chen, Y., Cheng, Y., Clune, T. L., Del Genio, A., Fainchtein,
1283 R., Faluvegi, G., Hansen, J. E., Healy, R. J., Kiang, N. Y., Koch, D., Lacis, A. A., LeGrande, A. N.,
1284 Lerner, J., Lo, K. K., Matthews, E. E., Menon, S., Miller, R. L., Oinas, V., Olosio, A. O., Perlwitz, J.
1285 P., Puma, M. J., Putman, W. M., Rind, D., Romanou, A., Sato, M., Shindell, D. T., Sun, S., Syed,
1286 R. A., Tausnev, N., Tsigaridis, K., Unger, N., Voulgarakis, A., Yao, M.-S., and Zhang, J.:
1287 Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive, *J. Adv.*
1288 *Model. Earth Syst.*, 6, 141–184, <https://doi.org/10.1002/2013MS000265>, 2014.

1289 Sepulchre, P., Caubel, A., Ladant, J.-B., Bopp, L., Boucher, O., Braconnot, P., Brockmann, P.,
1290 Cozic, A., Dufresne, J.-L., Estella-Perez, V., Foujols, M.-A., Ghattas, J., Hauglustaine, D.,
1291 Hourdin, F., Kageyama, M., Khodri, M., Krinner, G., Levavasseur, G., Marchand, M., Ottlé, C.,
1292 Peylin, P., Servonnat, J., Swingedouw, D., Szopa, S., Tardif, D., Terray, P., Vuichard, N., and
1293 Guilyardi, E.: IPSL-CM5A2 – an Earth system model designed for multi-millennial climate
1294 simulations, *Geosci. Model Dev.*, 13, 3011–3053, <https://doi.org/10.5194/gmd-13-3011-2020>,
1295 2020.

1296 Sueyoshi, T., Ohgaito, R., Yamamoto, A., Chikamoto, M. O., Hajima, T., Okajima, H., Yoshimori,
1297 M., Abe, M., O’ishi, R., Watanabe, S., and Abe-Ouchi, A.: Set-up of the PMIP3 paleoclimate
1298 experiments conducted using an Earth system model, MIROC-ESM, *Geosci. Model Dev.*, 6,
1299 819–836, <https://doi.org/10.5194/gmd-6-819-2013>, 2013.

1300 Sénési, S., Richon, J., Franchistéguy, L., Tyteca, S., Moine, M.-P., Voldoire, A., Sanchez-Gomez,
1301 E., Salas y Mélia, D., Decharme, B., Cassou, C., Valcke, S., Beau, I., Alias, A., Chevallier, M.,
1302 Déqué, M., Deshayes, J., Douville, H., Madec, G., Maisonnave, E., Planton, S., Saint-Martin, D.,
1303 Szopa, S., Alkama, R., Belamari, S., Braun, A., Coquart, L., and Chauvin, F.: CNRM-CM5 model
1304 output prepared for CMIP5 lgm, served by ESGF, World Data Center for Climate (WDCC) at
1305 DKRZ, doi: 10.1594/WDCC/CMIP5.CEC5lg, 2014.

1306

1307 S  n  si, S., Richon, J., Franchist  guy, L., Tyteca, S., Moine, M.-P., Voldoire, A., Sanchez-Gomez,
 1308 E., Salas y M  lia, D., Decharme, B., Cassou, C., Valcke, S., Beau, I., Alias, A., Chevallier, M.,
 1309 D  qu  , M., Deshayes, J., Douville, H., Madec, G., Maisonnave, E., Planton, S., Saint-Martin, D.,
 1310 Szopa, S., Alkama, R., Belamari, S., Braun, A., Coquart, L., and Chauvin, F.: CNRM-CM5 model
 1311 output prepared for CMIP5 piControl, served by ESGF, World Data Center for Climate (WDCC)
 1312 at DKRZ, doi: 10.1594/WDCC/CMIP5.CEC5pc, 2014.

1313 Tierney, J. E., Zhu, J., King, J., Malevich, S. B., Hakim, G. J., and Poulsen, C. J.: Glacial cooling
 1314 and climate sensitivity revisited, *Nature*, 584, 569–573, [https://doi.org/10.1038/s41586-020-](https://doi.org/10.1038/s41586-020-2617-x)
 1315 2617-x, 2020.

1316 Ullman, D. J., LeGrande, A. N., Carlson, A. E., Anslow, F. S., and Licciardi, J. M.: Assessing the
 1317 impact of Laurentide Ice Sheet topography on glacial climate, *Clim. Past*, 10, 487–507,
 1318 <https://doi.org/10.5194/cp-10-487-2014>, 2014.

1319 Voldoire, A., Sanchez-Gomez, E., Salas y M  lia, D., Decharme, B., Cassou, C., S  n  si, S., Valcke,
 1320 S., Beau, I., Alias, A., Chevallier, M., D  qu  , M., Deshayes, J., Douville, H., Fernandez, E.,
 1321 Madec, G., Maisonnave, E., Moine, M.-P., Planton, S., Saint-Martin, D., Szopa, S., Tyteca, S.,
 1322 Alkama, R., Belamari, S., Braun, A., Coquart, L., and Chauvin, F.: The CNRM-CM5.1 global
 1323 climate model: description and basic evaluation, *Clim. Dyn.*, 40, 2091–2121,
 1324 <https://doi.org/10.1007/s00382-011-1259-y>, 2013.

1325 Weller, E., and Cai, W.: Realism of the Indian Ocean Dipole in CMIP5 models: The implications
 1326 for climate projections, *J. Clim.*, 26, 6649–6659, <https://doi.org/10.1175/JCLI-D-12-00807.1>,
 1327 2013.

1328 Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Tanaka, T. Y.,
 1329 Shindo, E., Tsujino, H., Deushi, M., Mizuta, R., Yabu, S., Obata, A., Nakano, H., Koshiro, T., Ose,
 1330 T., and Kitoh, A.: A new global climate model of the Meteorological Research Institute: MRI-
 1331 CGCM3 – Model description and basic performance, *J. Meteorol. Soc. Jpn.*, 90, 23–64,
 1332 <https://doi.org/10.2151/jmsj.2012-A02>, 2012.

1333 Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Tanaka, T.,
 1334 Shindo, E., Tsujino, H., Deushi, M., Mizuta, R., Yabu, S., Obata, A., Nakano, H., Koshiro, T., Ose,
 1335 T., and Kitoh, A.: MRI-CGCM3 model output prepared for CMIP5 lgm, served by ESGF, World
 1336 Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.MRMClg, 2015.

1337 Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Tanaka, T.,
 1338 Shindo, E., Tsujino, H., Deushi, M., Mizuta, R., Yabu, S., Obata, A., Nakano, H., Koshiro, T., Ose,
 1339 T., and Kitoh, A.: MRI-CGCM3 model output prepared for CMIP5 piControl, served by ESGF,
 1340 World Data Center for Climate (WDCC) at DKRZ, doi: 10.1594/WDCC/CMIP5.MRMCpc, 2015.

1341 Zhang, Y., Du, Y., and Qu, T.: A sea surface salinity dipole mode in the tropical Indian Ocean,
 1342 *Clim. Dyn.*, 47, 2573–2585, <https://doi.org/10.1007/s00382-016-2984-z>, 2016.

Supplementary Materials for:

Using ocean surface paleo-density to evaluate PMIP3 and PMIP4 Last Glacial Maximum climate simulations.

Héloïse Barathieu ¹, Thibaut Caley ¹, Masa Kageyama ², Didier Swingedouw ¹, Pascale Braconnot ²

¹Univ. Bordeaux, CNRS, Bordeaux INP, EPOC, UMR 5805, 33600 Pessac, France

²Laboratoire des Sciences du Climat et de l'Environnement – Institut Pierre-Simon Laplace (LSCE-IPSL), UMR CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France.

Correspondence to: Héloïse Barathieu (*heloise.barathieu@u-bordeaux.fr*) or Thibaut Caley (*thibaut.caley@u-bordeaux.fr*)

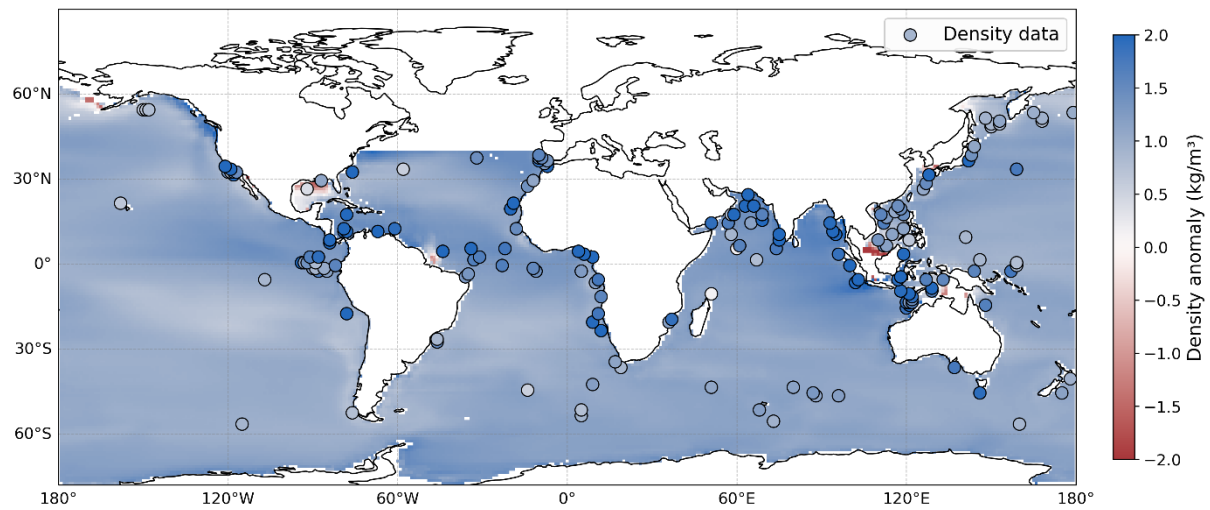
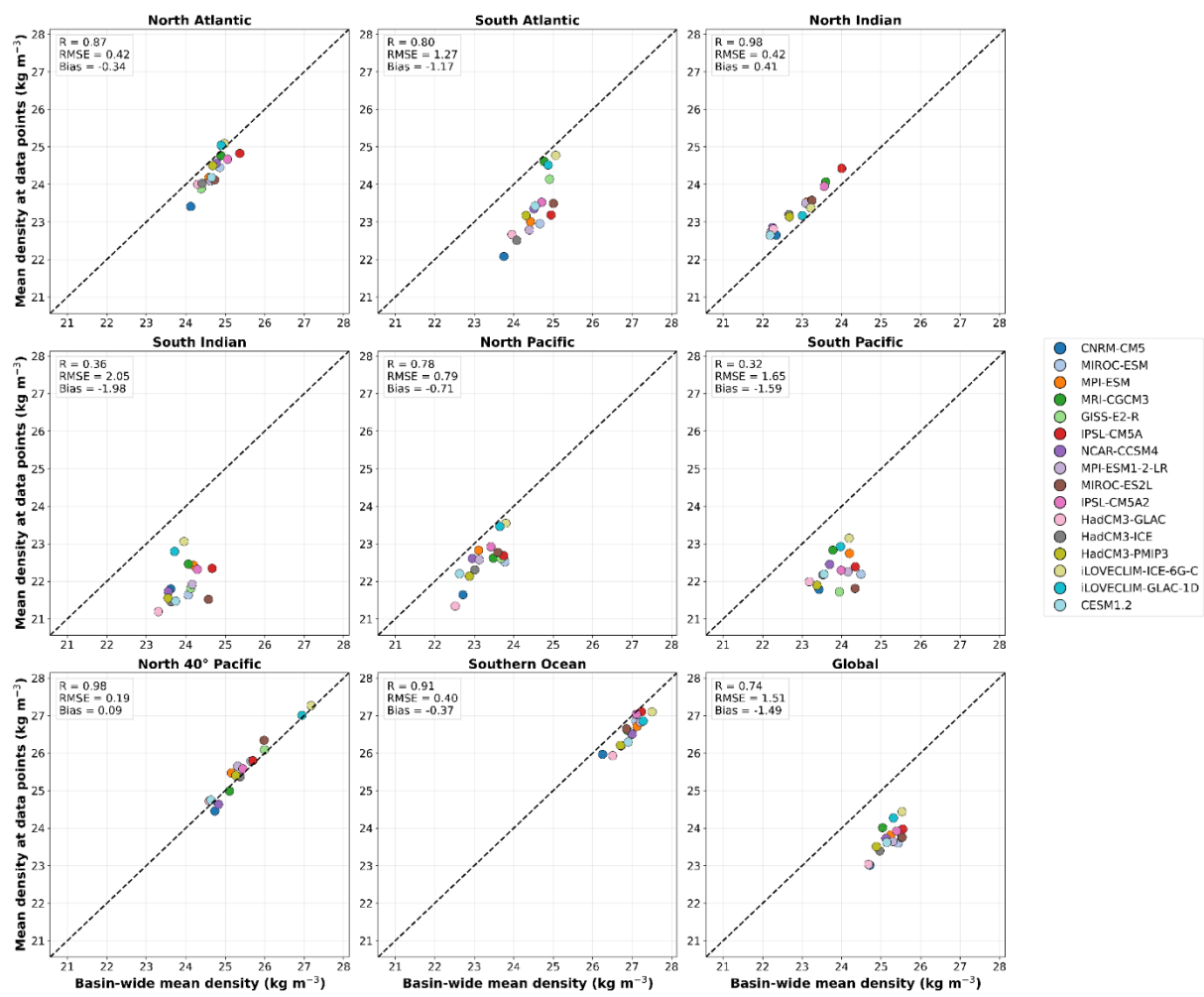


Figure S1: Absolute surface density (kg/m^3) anomaly map (LGM - piControl). The dots represent the surface density anomaly database reconstructions (Caley et al., 2026) and the background map is the anomaly mean for all model simulations in the study.

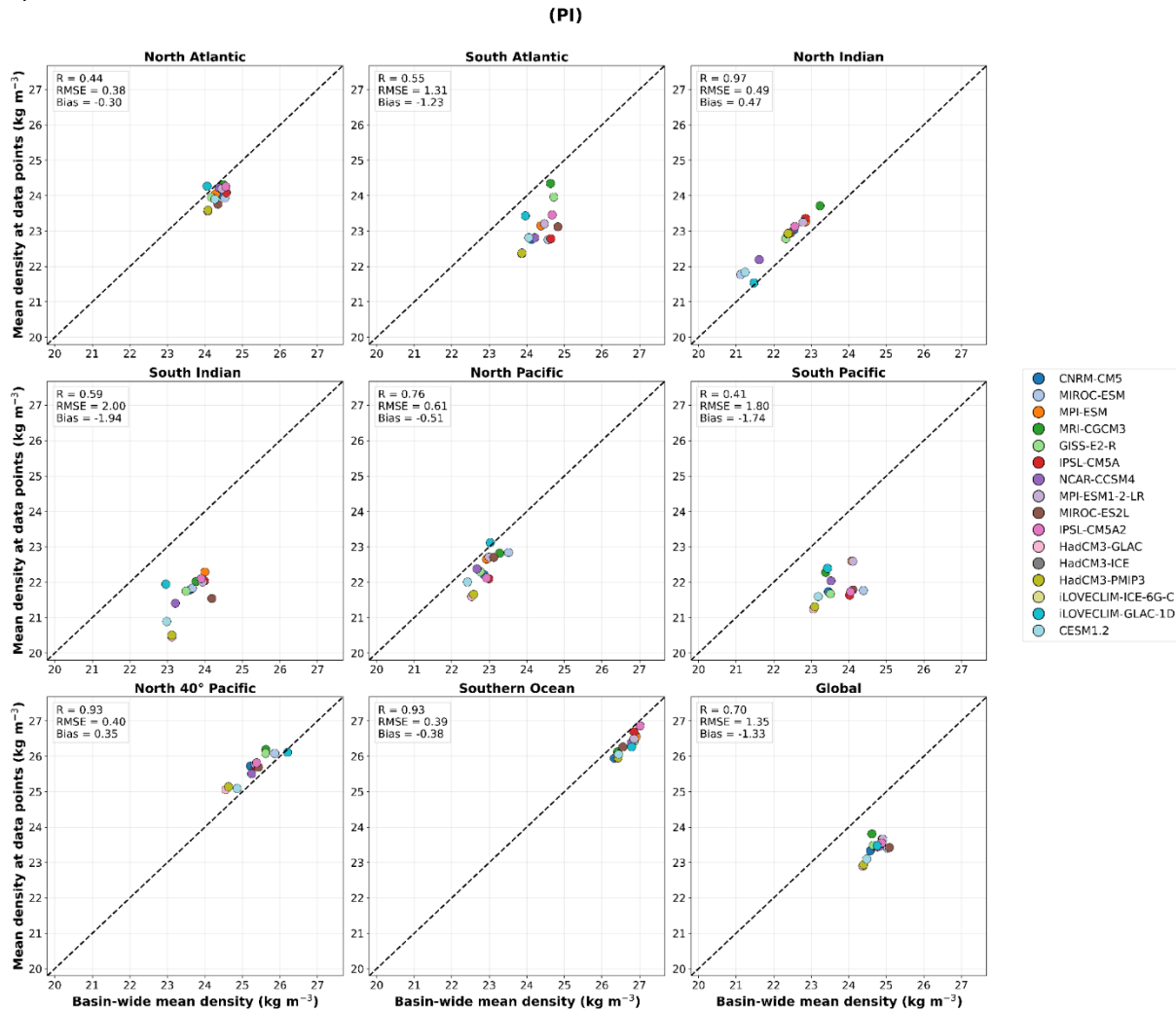
33 a)

(LGM)



34

35 b)



36

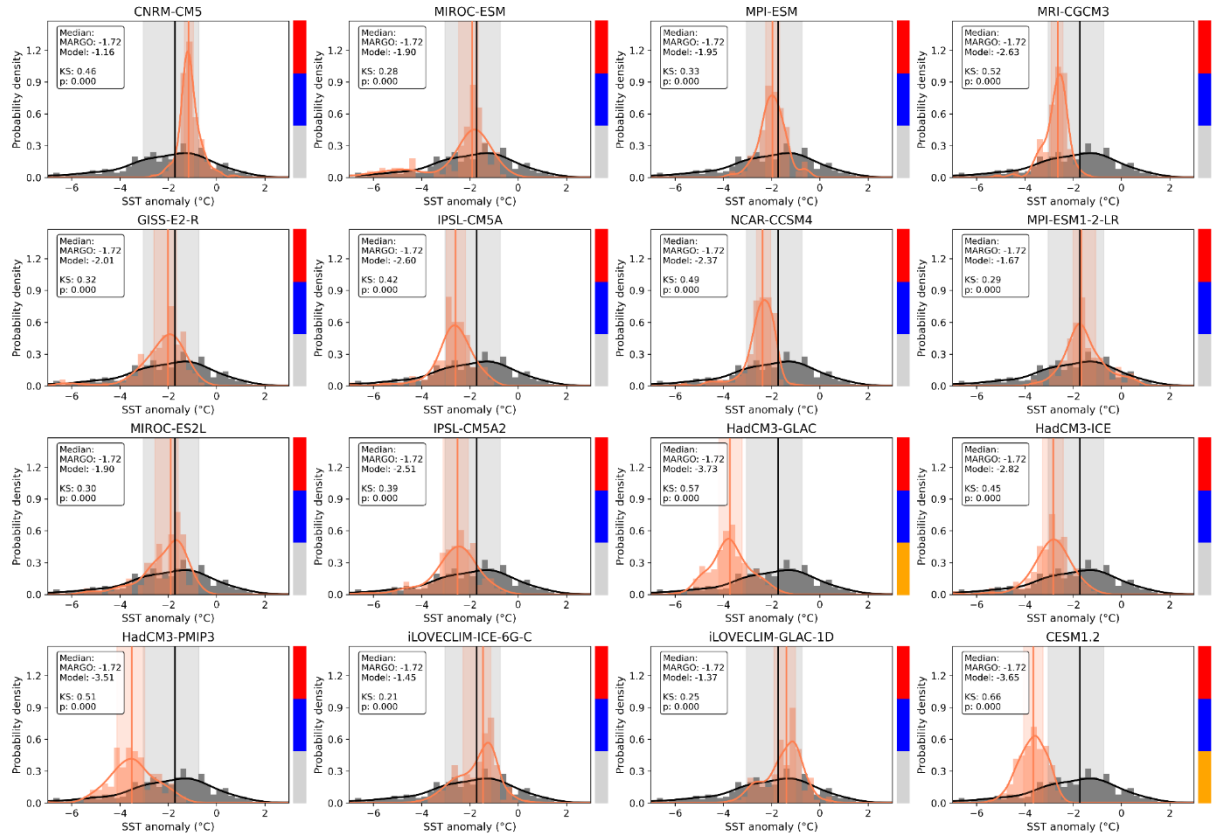
37 **Figure S2:** Pseudo-proxy test of the spatial representativeness of the proxy network for the LGM (a) and PI (b) ocean surface
 38 density. For each ocean basin and for the global ocean, the mean density simulated by each climate model over all grid cells
 39 within the basin (x-axis) is compared with the mean density obtained by sampling the same model only at the locations where
 40 proxy reconstructions are available (y-axis). Each point represents one model simulation, and the dashed line indicates the
 41 1:1 relationship. This pseudo-proxy approach does not involve proxy system modelling. Instead, it evaluates whether the
 42 spatial distribution of proxy sites provides an unbiased estimate of the basin-scale density signal. Statistics shown in each
 43 panel include the Pearson correlation coefficient (R), bias, and root mean square error (RMSE) between basin-wide model
 44 means and model values sampled at proxy locations. A close agreement with the 1:1 line indicates that the proxy network
 45 accurately captures the large-scale basin mean, whereas systematic departures reflect spatial sampling biases arising from
 46 the uneven distribution of proxy records. The “Global” panel includes all proxy locations from the individual basins shown.

47

48

49

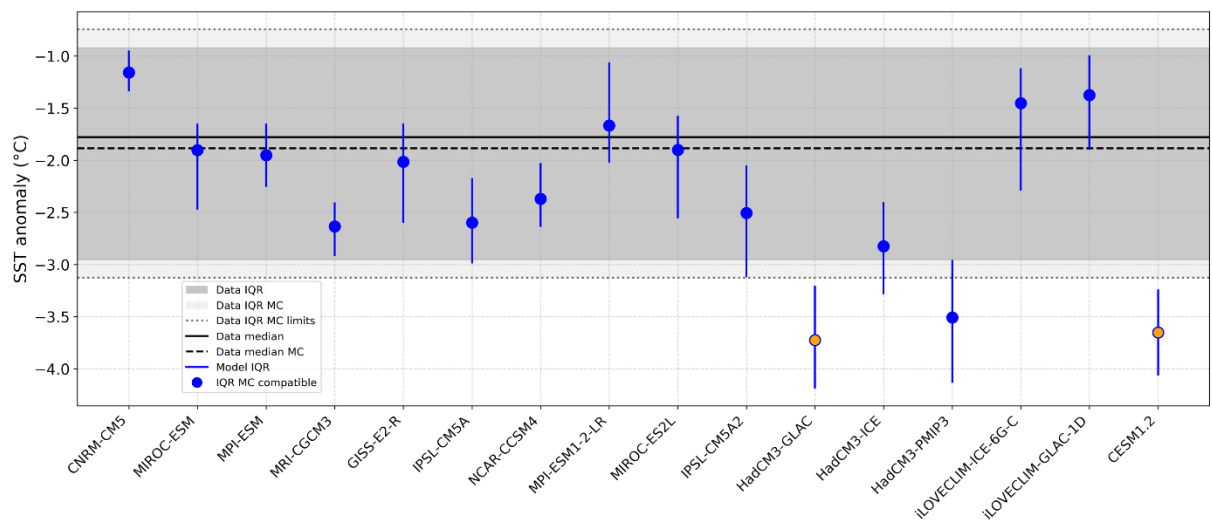
a)



50

51

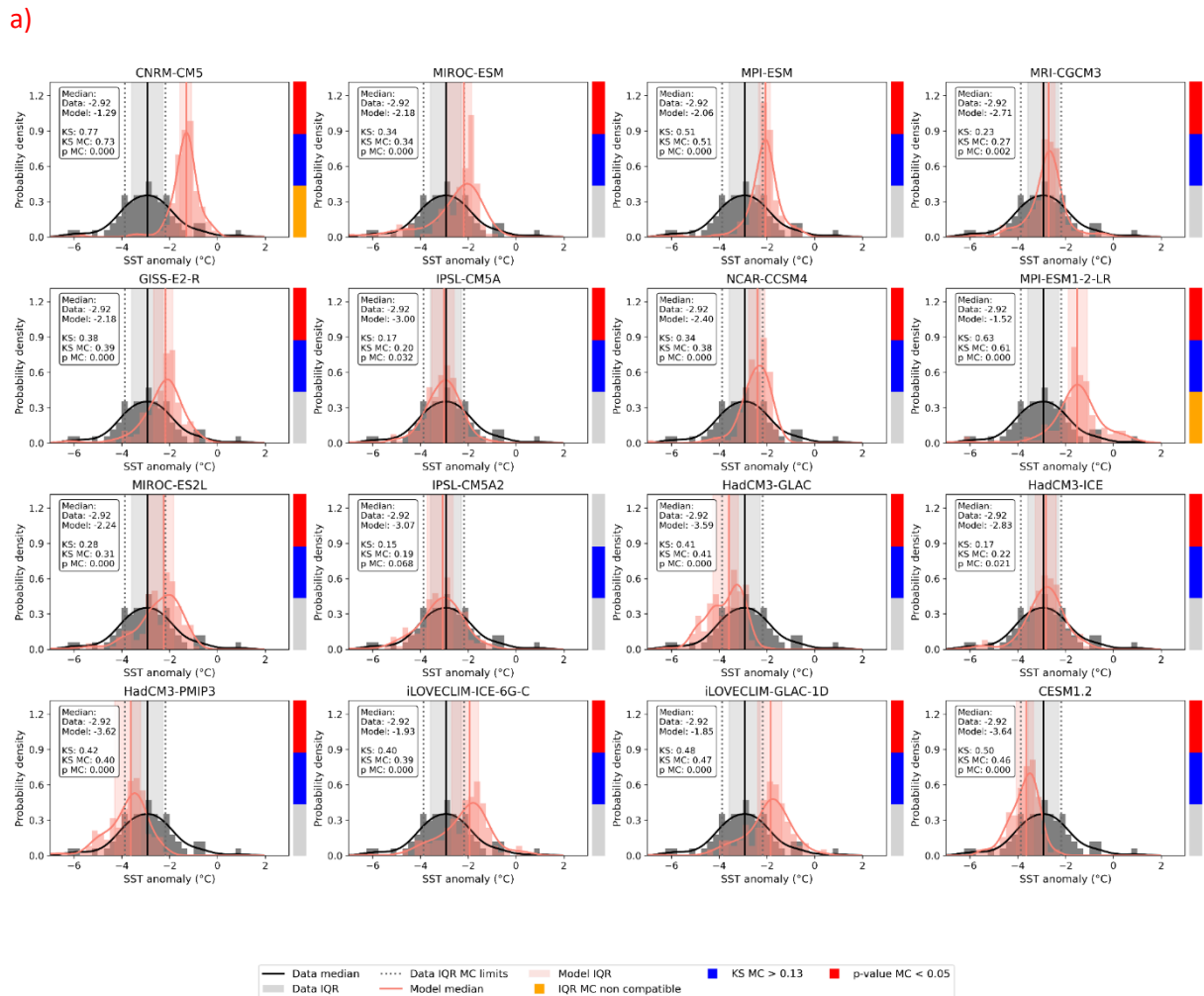
b)



52

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

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



b)

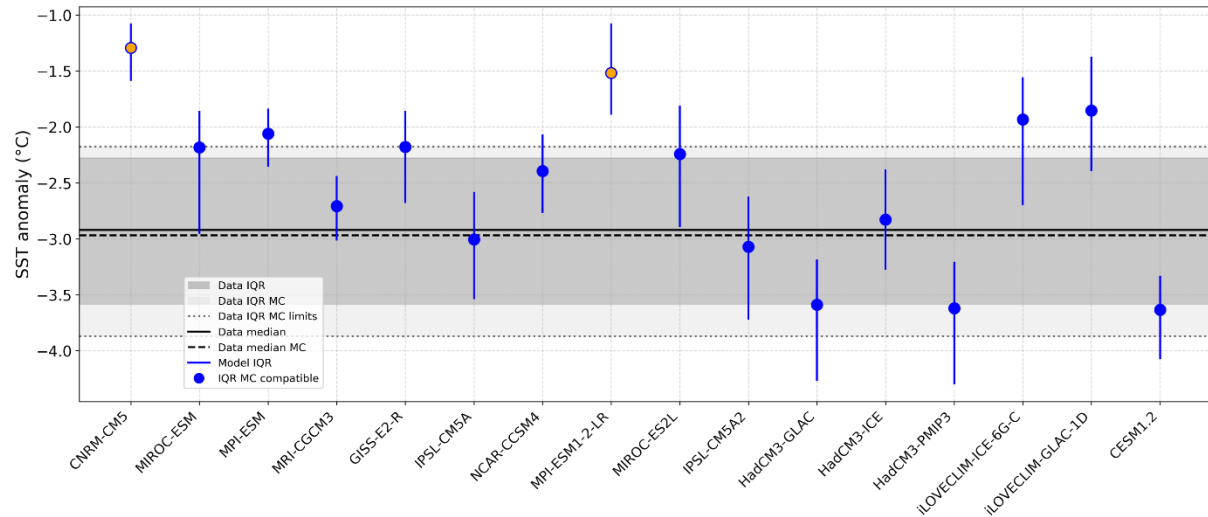
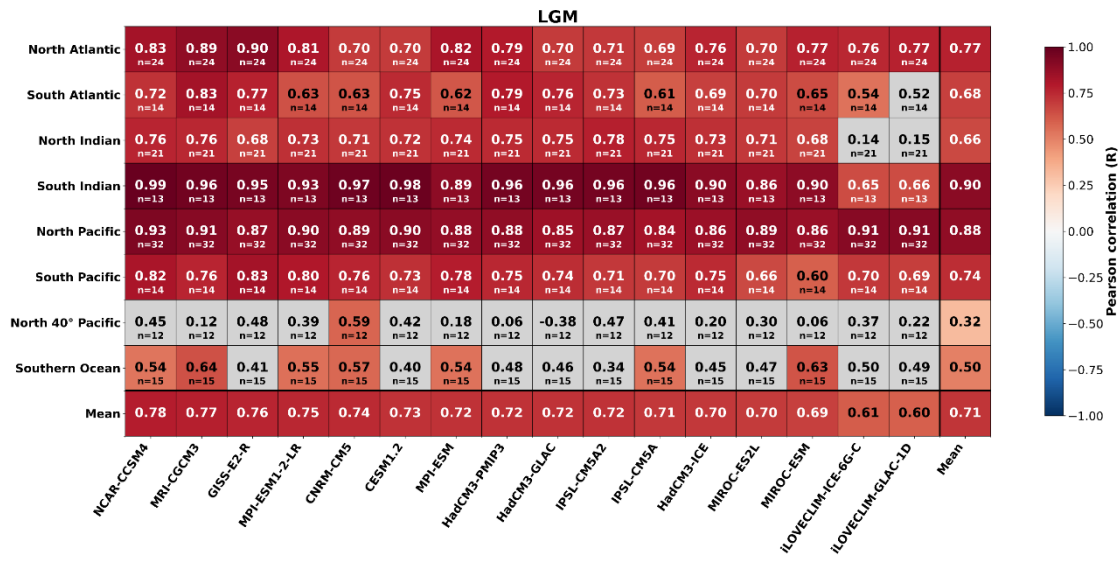


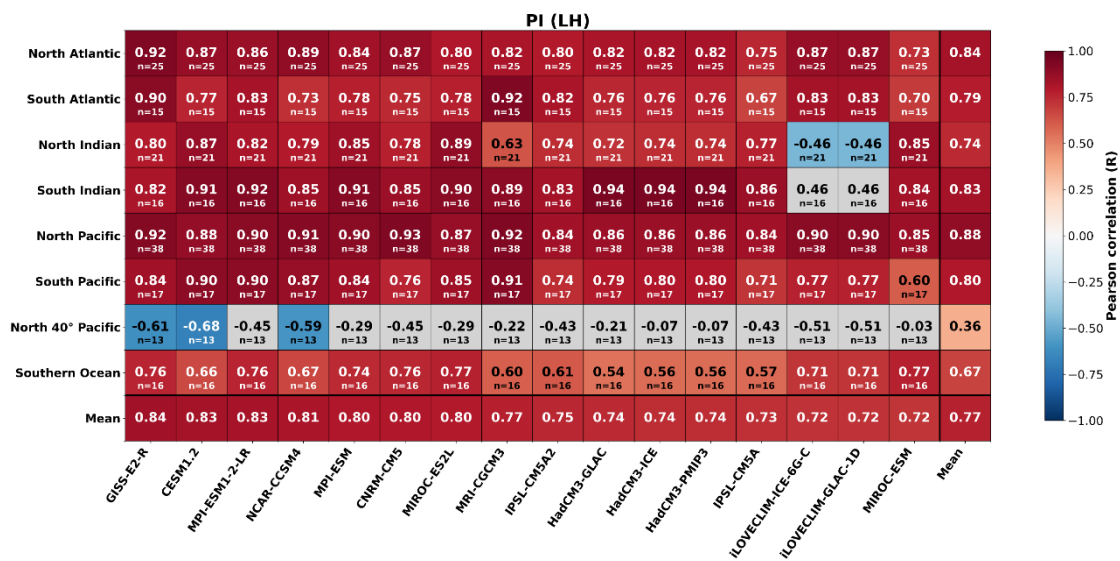
Figure S4: (a) Distribution histograms of SST anomalies (LGM–PI, °C) for each ocean basin. SST reconstructions from the Tierney database (Tierney et al., 2020) are shown in black, and model simulations in orange. Kernel Density Estimates (KDEs) illustrate the central tendency and shape of the distributions.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

105 a)



106

107 b)



108

109 **Figure S5:** Pearson correlation coefficients (R) between reconstructed seawater density derived from proxy-based estimates
 110 and density fields simulated by climate models for **a)** the Last Glacial Maximum (LGM) and **b)** pre-industrial (PI) climates.
 111 Correlations were evaluated independently within eight ocean basins (see Figure A1) using grid cells common to both
 112 reconstructed and simulated fields after interpolation onto a common spatial grid. The number of valid grid cells used in each
 113 comparison (n) is indicated within each cell. Models are ordered according to their weighted mean absolute correlation ($|R|$),
 114 calculated using the number of valid grid cells as weights. Colors represent the Pearson correlation coefficient, with blue
 115 indicating negative correlations and red indicating positive correlations. Light-gray cells denote correlations that are not
 116 statistically significant ($p \geq 0.05$). The “Mean” row and column show weighted averages of $|R|$ across basins and models,
 117 respectively, and therefore provide a measure of overall pattern agreement independent of correlation sign. The lower-right
 118 value corresponds to the global weighted mean absolute correlation.

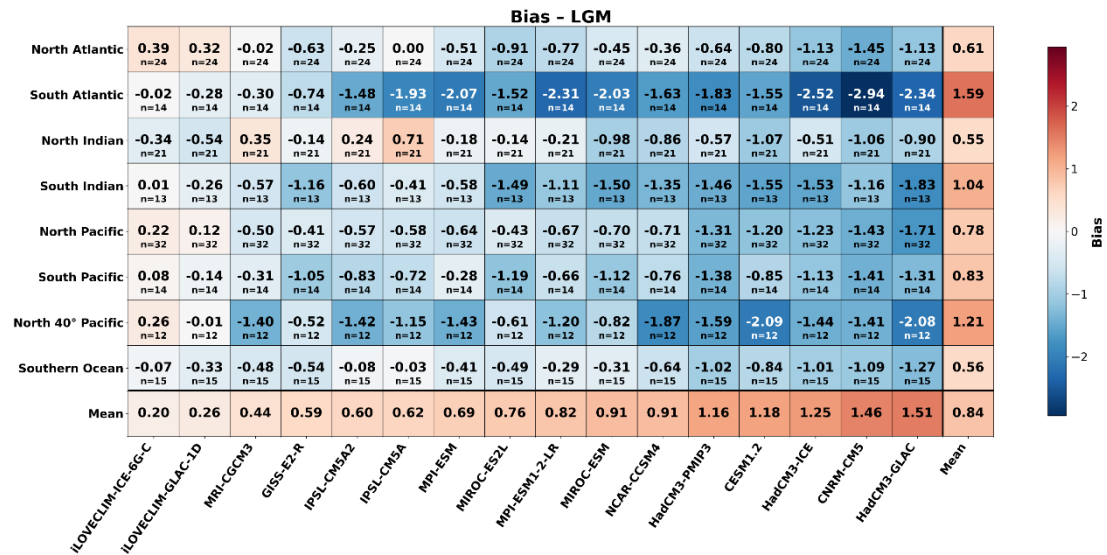
119

120

121

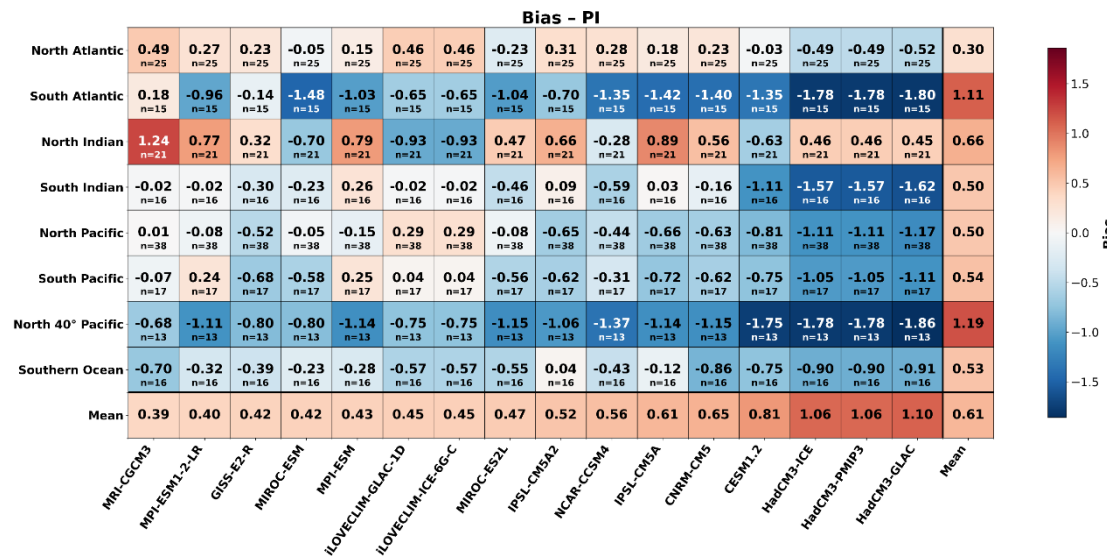
122

123 a)



124

125 b)



126

127 **Figure S6:** Weighted basin-scale mean density bias between simulated and reconstructed seawater density for the eight
 128 ocean basins during **a)** the Last Glacial Maximum (LGM) and **b)** pre-industrial (PI). Rows correspond to ocean basins and
 129 columns to climate model simulations. Individual cells show the mean bias (model – reconstruction) computed from grid
 130 points common to all simulations and reconstructions within each basin. The number of valid grid points (n) is indicated below
 131 each value. Models are ranked from left to right according to their weighted mean absolute bias, with lower values indicating
 132 better agreement with the reconstructions. The final column and row (“Mean”) show weighted averages based on the
 133 number of valid grid points in each basin. For these averages, absolute bias values are used to summarize model performance
 134 independently of the sign of the error. Positive values indicate simulated densities that are higher than reconstructed
 135 densities, whereas negative values indicate underestimation. The colour scale is centred on zero, with blue colours
 136 representing negative biases and red colours representing positive biases.

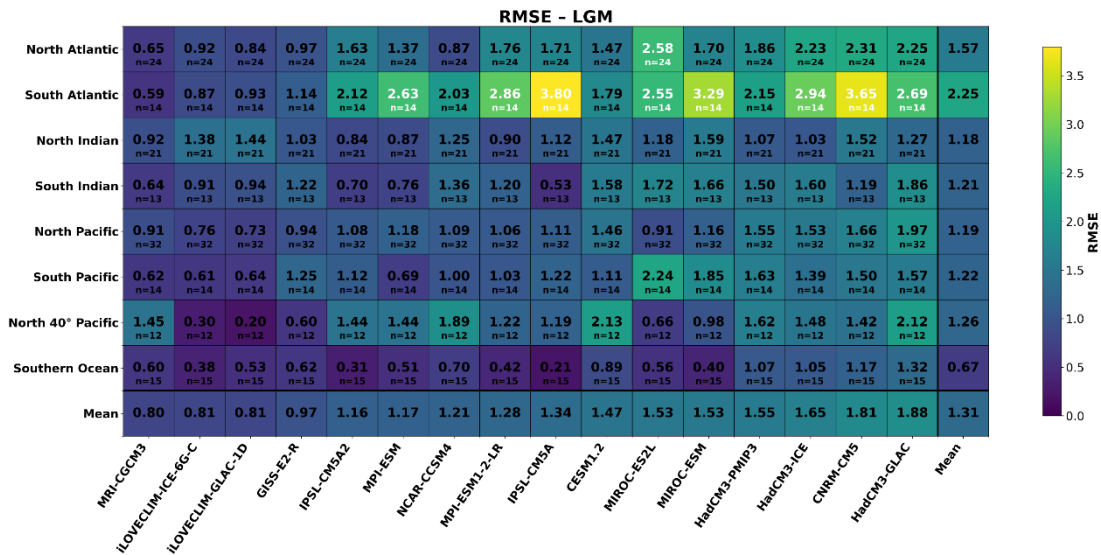
137

138

139

140

a)



b)

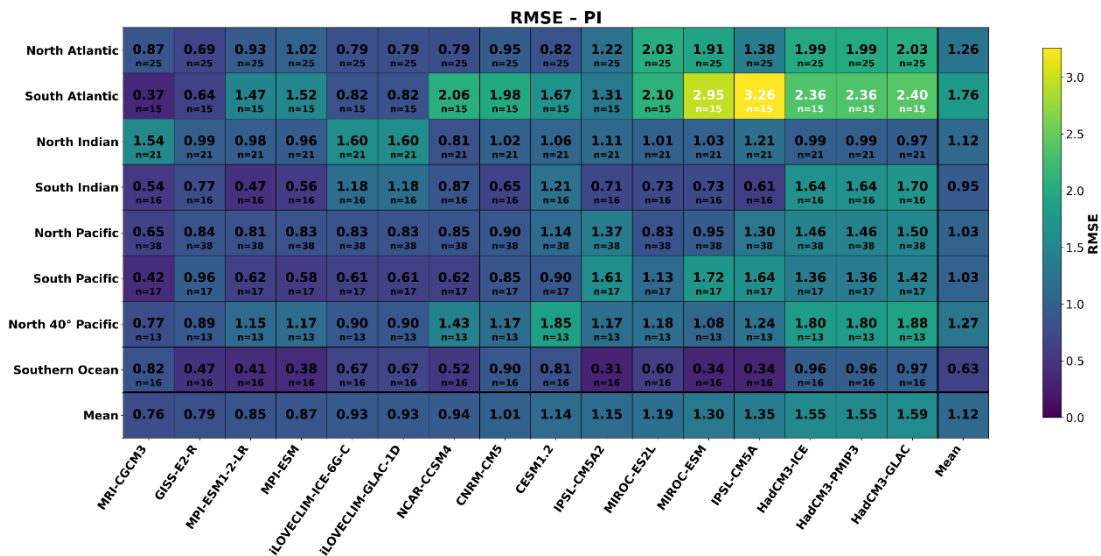


Figure S7: Weighted basin-scale root mean square error (RMSE) between simulated and reconstructed seawater density for the eight ocean basins during **a)** the Last Glacial Maximum (LGM) and **b)** pre-industrial (PI). Rows correspond to ocean basins and columns to climate model simulations. Individual cells show the RMSE calculated from grid points common to all simulations and reconstructions within each basin. The number of valid grid points (n) is indicated below each value. Models are ranked from left to right according to their weighted mean RMSE, with lower values indicating better agreement with the reconstructions. The final column and row ("Mean") show weighted averages based on the number of valid grid points in each basin. RMSE values are always positive and provide a measure of the overall magnitude of model–data mismatch, combining both systematic and spatially varying errors. Lower RMSE values indicate a closer agreement between simulated and reconstructed density patterns. Colours range from low errors (dark purple) to high errors (yellow).