



Supplement of

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

Héloïse Barathieu et al.

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

The copyright of individual parts of the supplement might differ from the article licence.

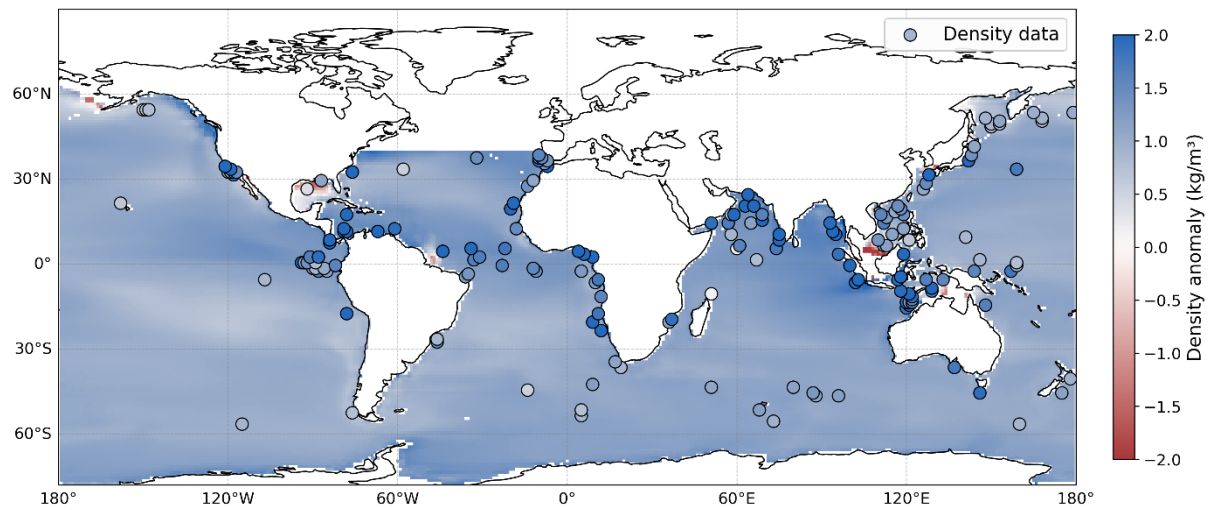
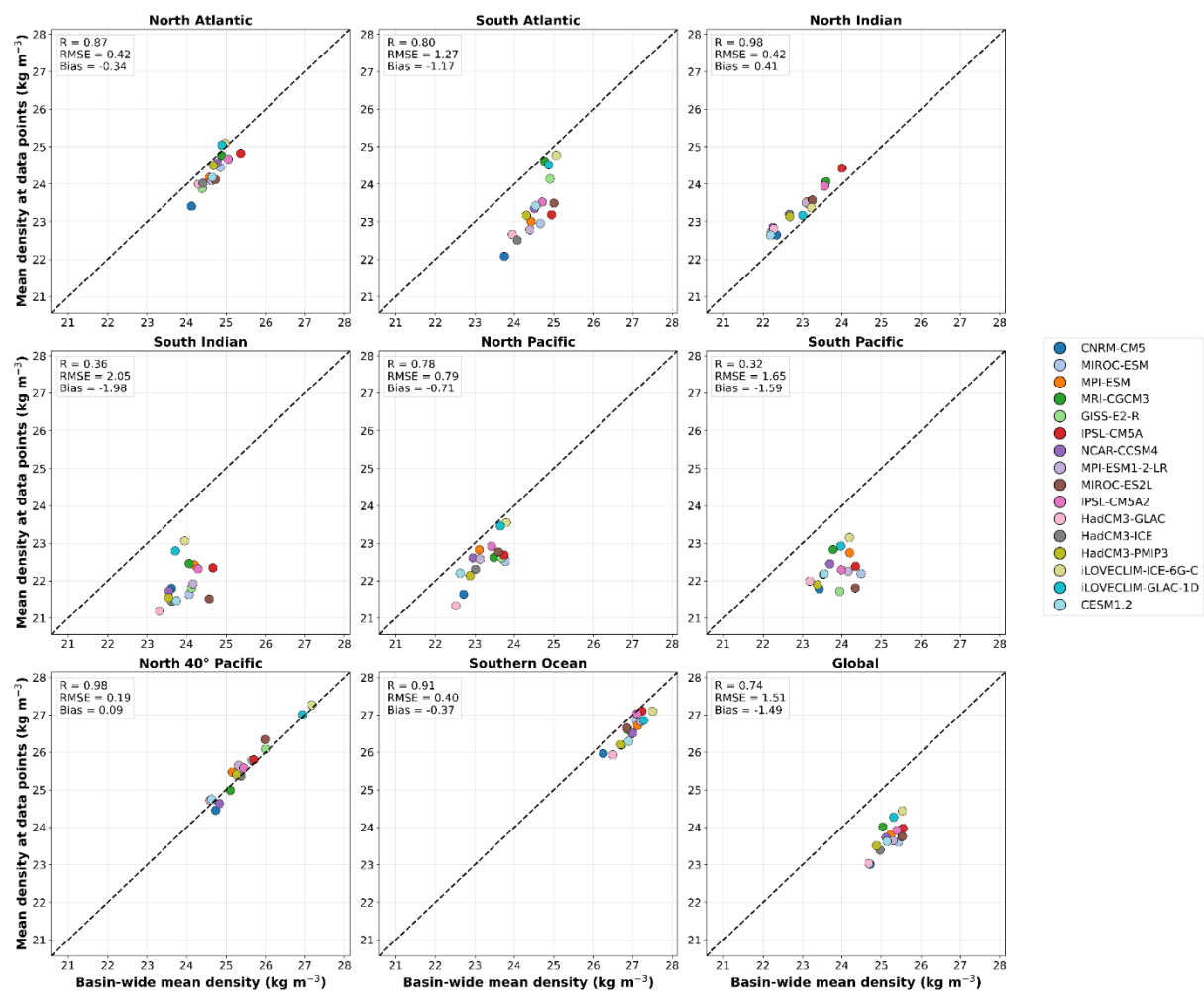


Figure S1: Absolute surface density (kg/m^3) anomaly map (LGM - piControl). The dots represent the surface density anomaly database reconstructions (Caley et al., 2026) and the background map is the anomaly mean for all model simulations in the study.

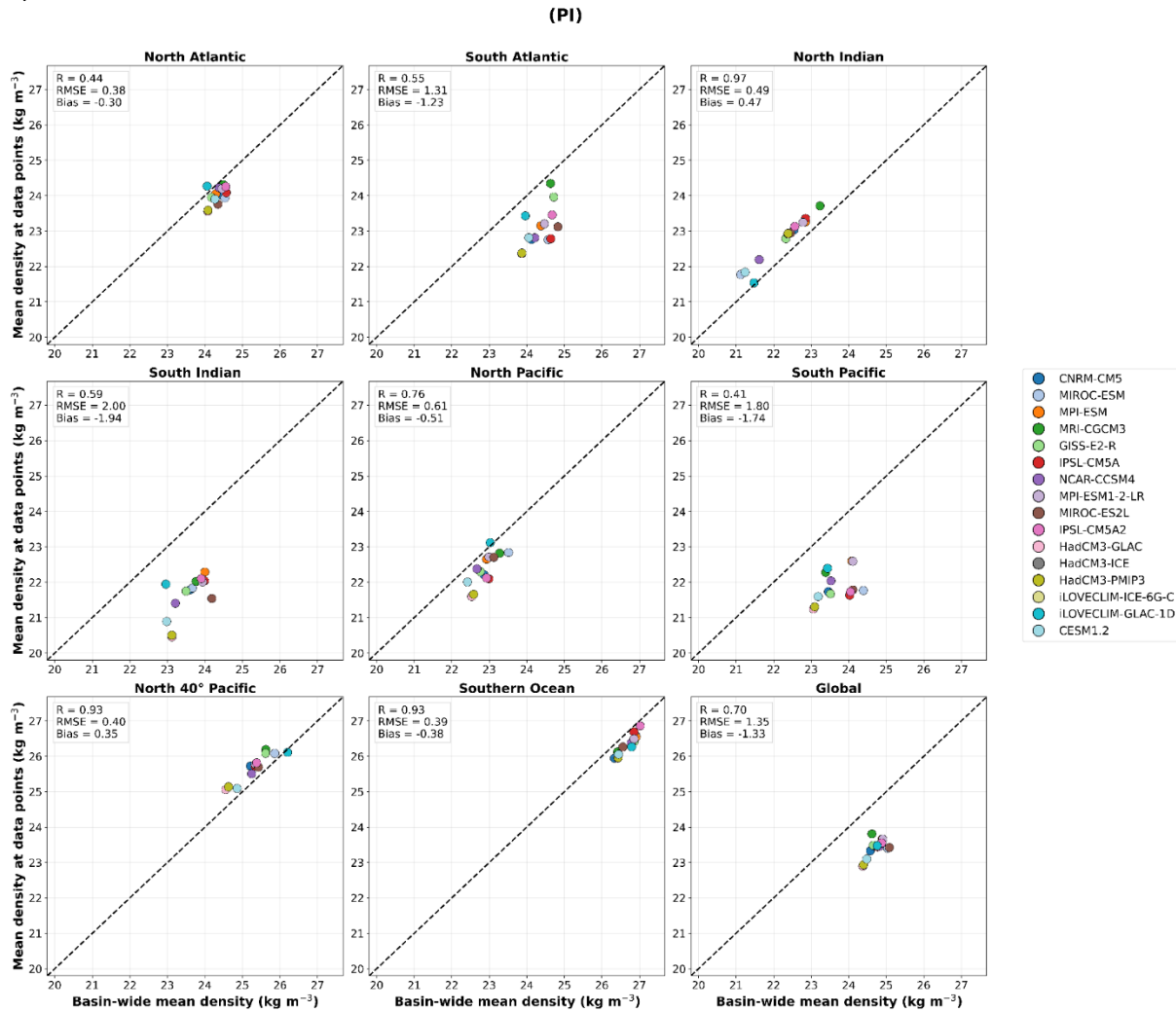
33 a)

(LGM)



34

35 b)



36

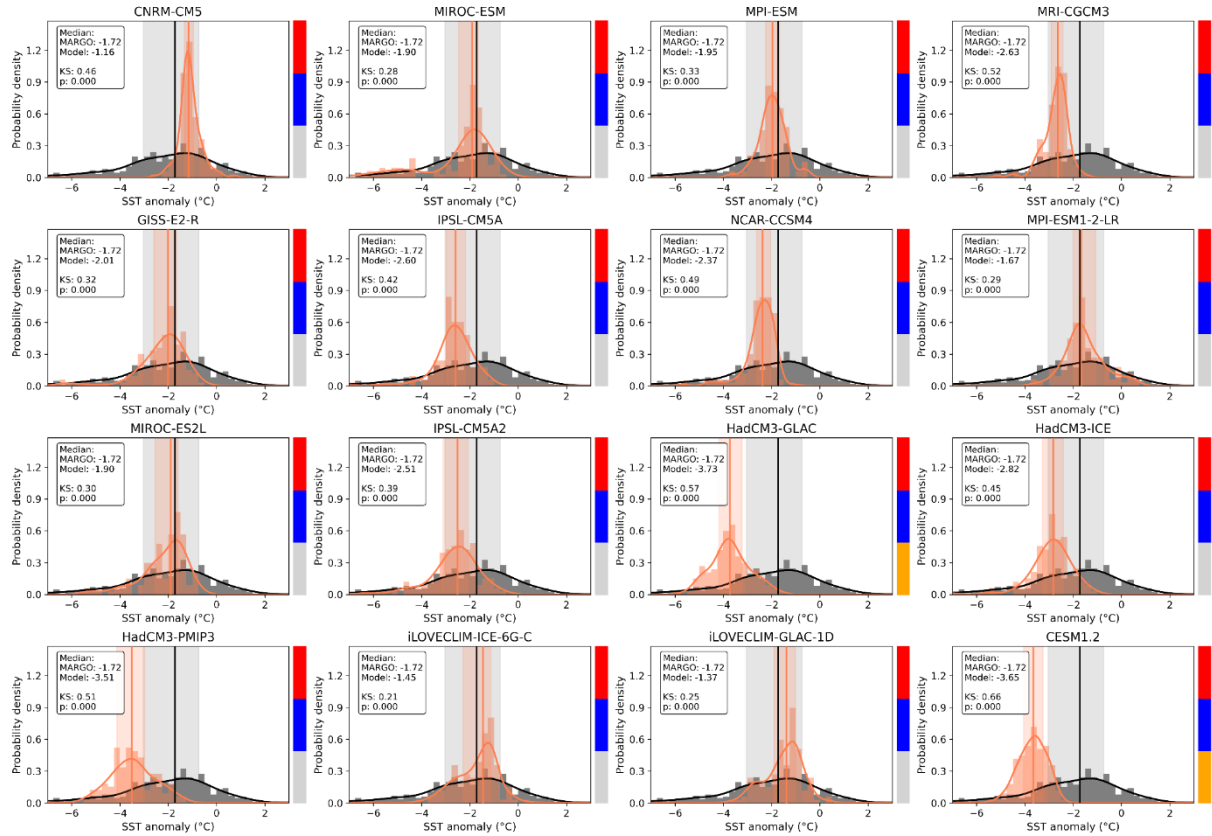
37 **Figure S2:** Pseudo-proxy test of the spatial representativeness of the proxy network for the LGM (a) and PI (b) ocean surface
 38 density. For each ocean basin and for the global ocean, the mean density simulated by each climate model over all grid cells
 39 within the basin (x-axis) is compared 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

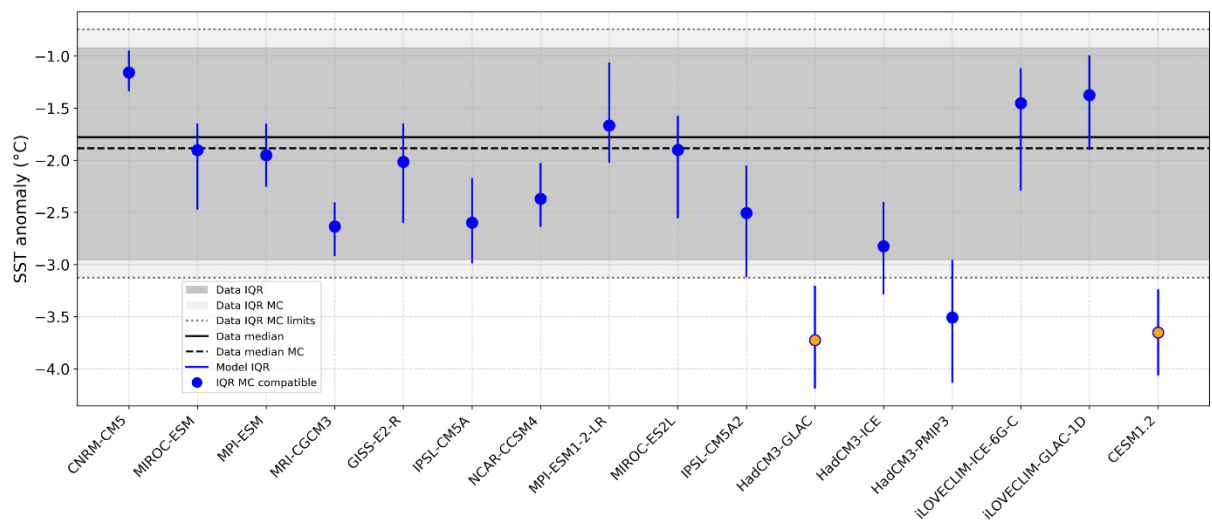
a)



50

51

b)

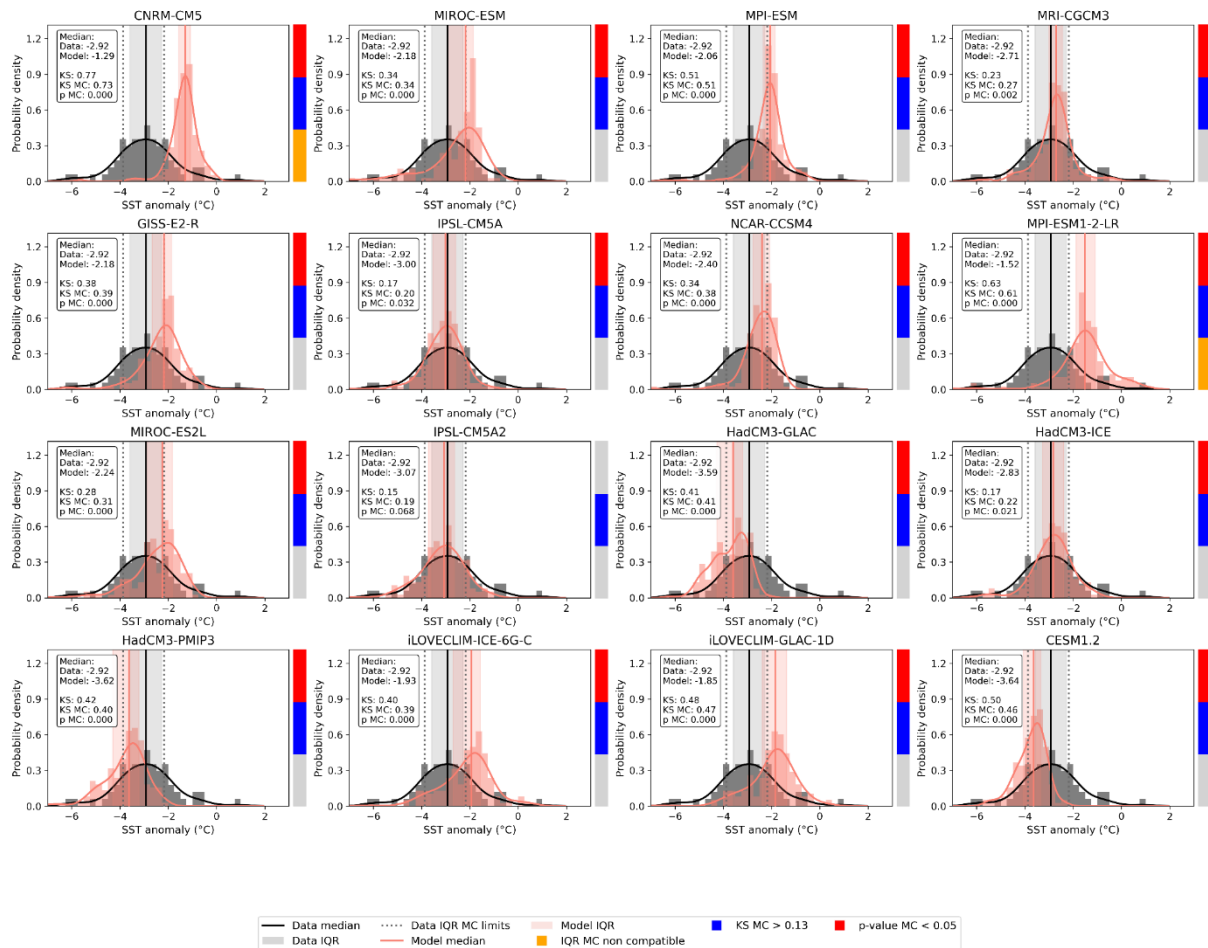


52

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.

a)



b)

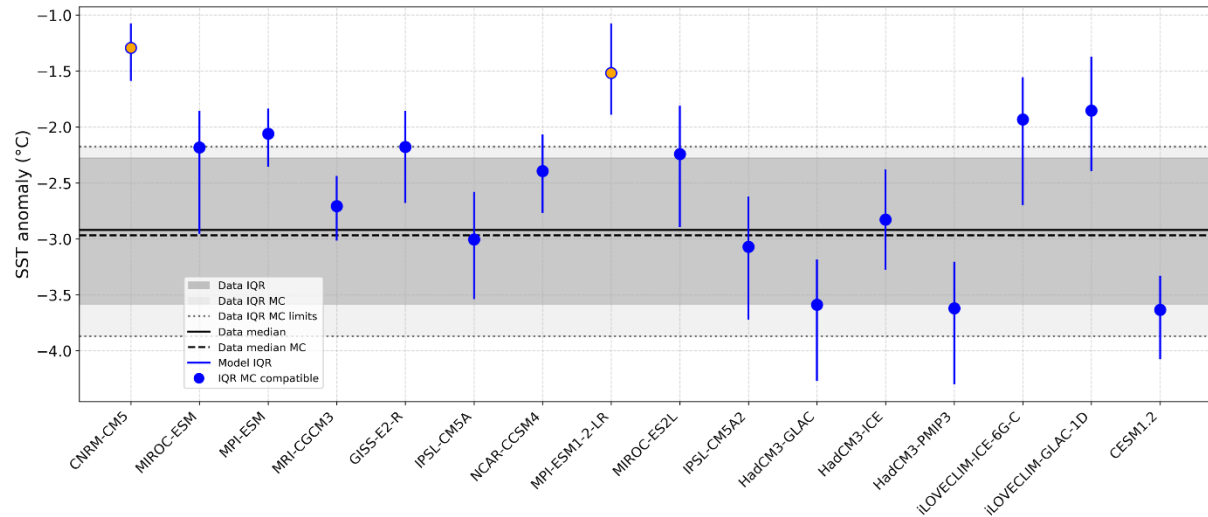
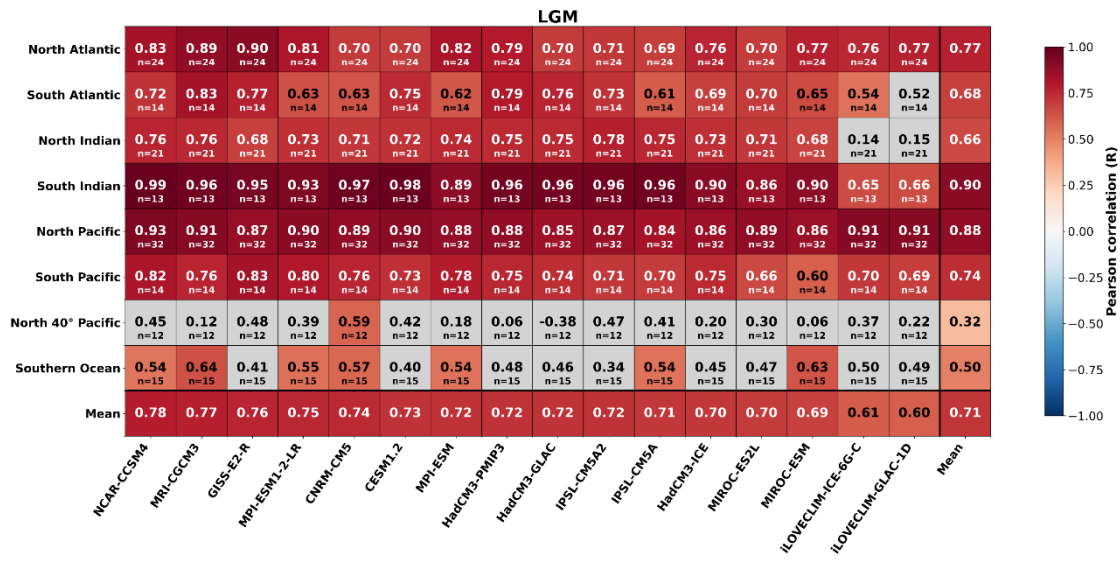


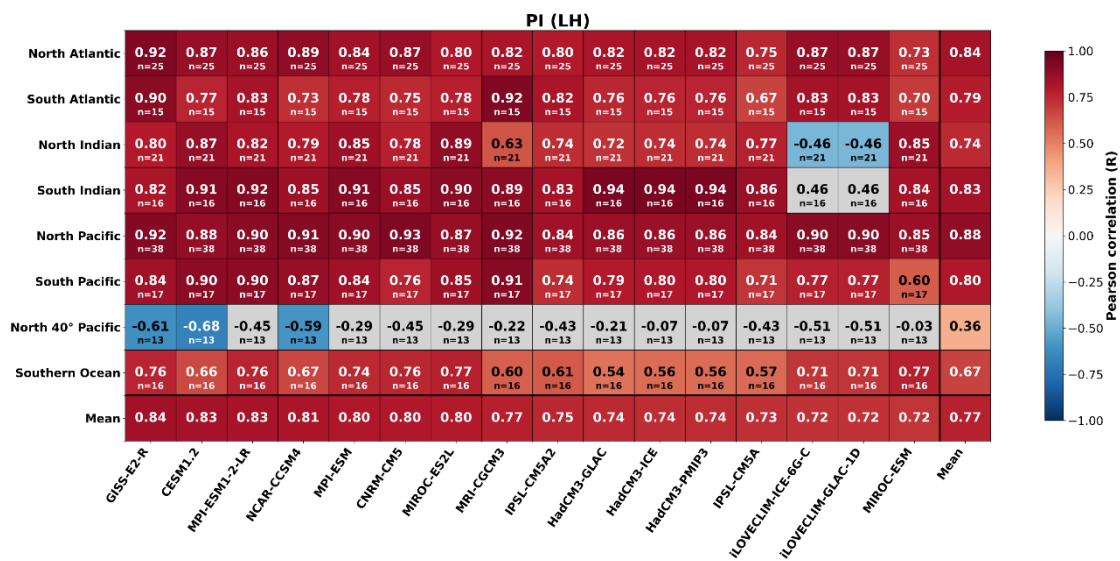
Figure S4: (a) Distribution histograms of SST anomalies (LGM–PI, °C) for each ocean basin. SST reconstructions from the Tierney database (Tierney et al., 2020) are shown in black, and model simulations in orange. Kernel Density Estimates (KDEs) illustrate the central tendency and shape of the distributions. 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.

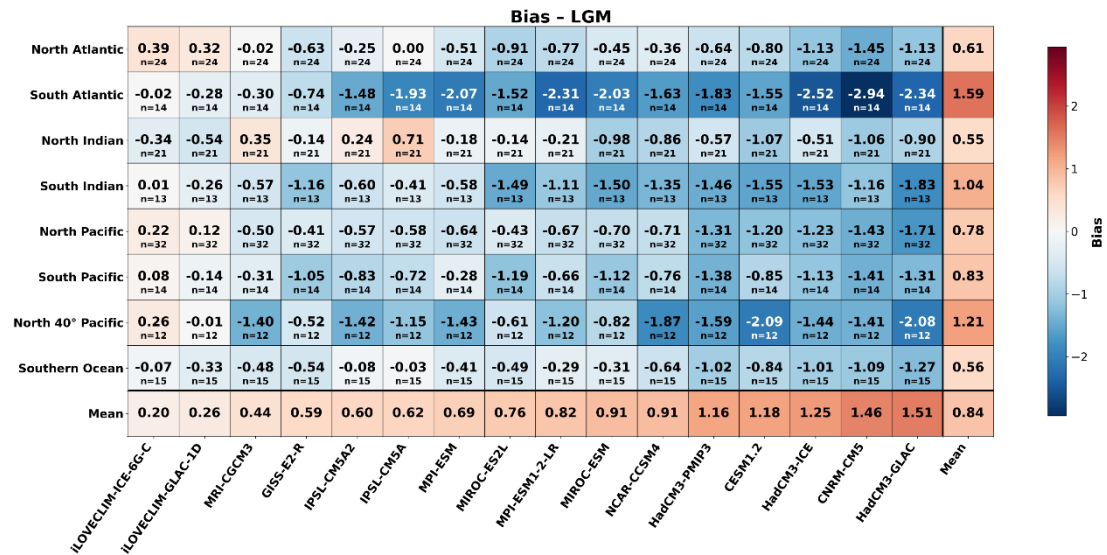
119

120

121

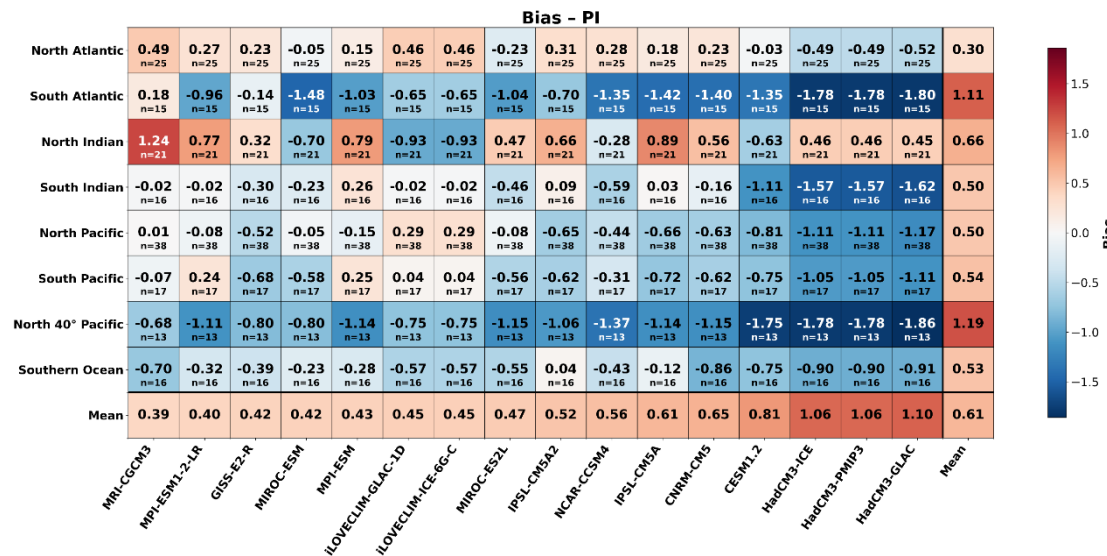
122

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.

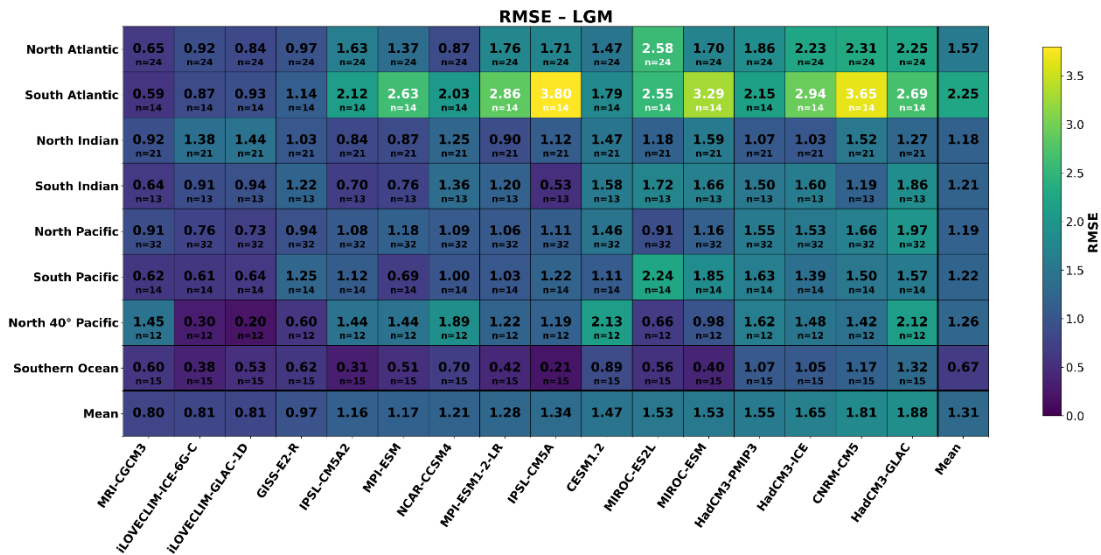
137

138

139

140

a)



b)

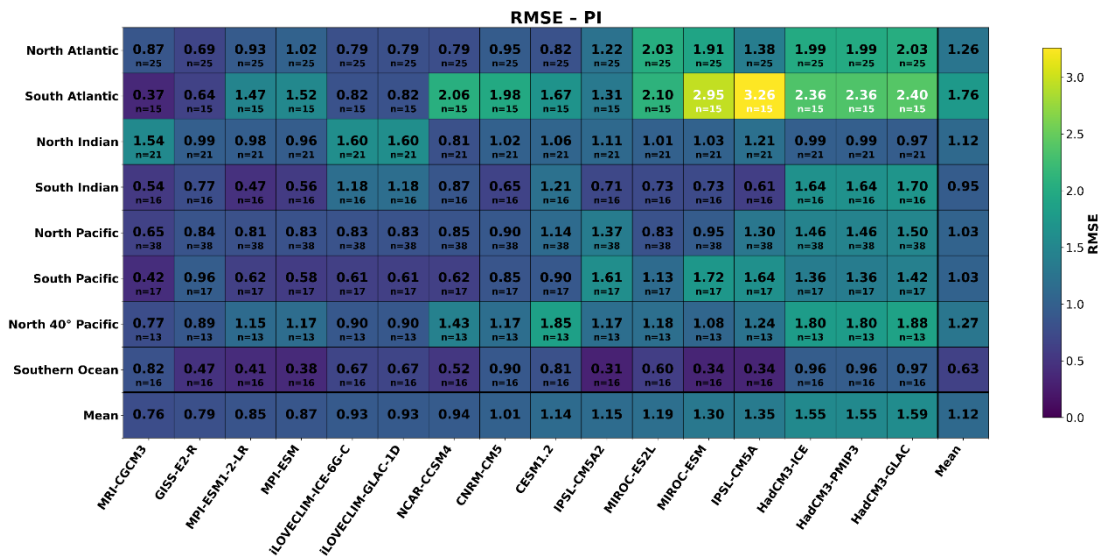


Figure S7: Weighted basin-scale root mean square error (RMSE) between simulated and reconstructed seawater density for the eight ocean basins during **a)** the Last Glacial Maximum (LGM) and **b)** pre-industrial (PI). Rows correspond to ocean basins and columns to climate model simulations. Individual 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).