We are grateful to both reviewers for their positive assessment of the revised manuscript and for their suggestions. We have addressed each of the minor comments raised, as detailed in our point-by-point replies below.

Our replies are displayed in blue. The line numbers mentioned in our replies refer to the marked-up version of the revised manuscript.

Report #1

Referee: Sam Sherriff-Tadano

L20: It might be better to add “past” before the word “changes.”

Reply: We have added “past” as suggested.

L121: It would be helpful to explain the advantages and limitations of using RMSE and Cohen’s κ in this paragraph so that readers can better understand why these two metrics are used and how they complement each other.

Reply: We have improved this paragraph as suggested (L121-130):

To quantitatively assess the proxy-model agreement, we use two complementary statistical metrics: root mean square error (RMSE) and Cohen’s κ with binary categories (“positive anomaly” vs. “negative anomaly”). The proxy sites are projected to the nearest grid cell in each model’s native grid prior to the analyses. RMSE is expressed in the units of the environmental variable and is therefore directly interpretable and comparable across models; however, it is dominated by amplitude errors, meaning that a model simulating little change can score better than one that reproduces the correct spatial pattern but excessive amplitude. Cohen’s κ with binary categories instead evaluates only the direction of change rather than amplitude, while also corrects for the agreement expected by chance. Its limitations are that sites falling close to the category threshold are weighted the same as unambiguous ones, and that its robustness is dependent on sample size. Therefore, combining these two metrics is usually more informative than either metric alone.

L149: It would be better to refer to Fig. S1 in this sentence.

Reply: Done, the sentence now reads:

We further analysed the seasonal cycle of MLD averaged over three deep-water formation regions as defined by Treguier et al. (2023): the Labrador Sea, Irminger Sea, and Greenland Sea (see also Fig. S1).

L159–160: It would be preferable to define the metrics used for identifying outliers and compare the number of such models across the different PMIP phases. Alternatively, the authors might want to remove this sentence entirely unless they feel it is essential to keep.

Reply: We have removed this sentence as suggested.

L188–190: I think these sentences could be rewritten in a more positive tone and perhaps relocated to the Methods (L121-)/Discussion section.

Reply: We have rewritten this paragraph (L196-205):

As models often simulate MLD anomalies of the opposite sign to the reconstructions, differences in RMSE between models largely reflect differences in simulated amplitude rather than in agreement with the reconstructions. Consequently, a low RMSE may reflect a weak simulated response rather than better agreement (Sect. 2). We therefore do not use RMSE to rank model performance. The RMSE remains informative in absolute terms: for every model, it exceeds the RMSEP of the reconstruction itself, indicating that the proxy-model disagreement is larger than can be attributed to reconstruction uncertainty alone. The standard deviation across the ensemble is comparable to the ensemble-mean signal of the MLD anomaly for PMIP4, but larger for PMIP3 (Fig. 5b, d).

L270–273: The points raised in these two sentences seem to overlap. It would be better to remove one of them to avoid redundancy.

Reply: We have modified this part as suggested, it now reads (L283-287):

In contrast, in areas that are ice-free in PI, the MLD generally increases in response to enhanced ocean heat loss under reduced winter insolation, with no accompanying change in sea ice (e.g., CESM2, GISS-E2.1G, MPI-ESM1.2-LR, MRI-ESM2.0, NorESM2-LM). The intensified deep convection reduces stratification, so conditions there remain unfavourable for local sea ice formation.

L304–306: If the statistical significance is questionable, it might be better to omit this sentence entirely to improve the flow of the text.

Reply: We have revised the second half of this sentence to a more neutral tone (L318-322):

In addition, the difference between HadGEM2-CC and HadGEM2-ES suggests that the implementation of tropospheric chemistry might mitigate the model's response to MH

orbital forcing (see Fig. 2). However, with only 35 years available for HadGEM2-CC, we cannot establish whether this difference exceeds internal variability, and we note this as a potential direction for future work.

L346–359: The fact that the impact of meltwater during the Mid-Holocene is difficult to discern in some models should not be viewed negatively. While non-freshwater factors may also play a role, as the authors state in the subsequent paragraph, model resolutions remain coarse, and there are substantial uncertainties regarding how meltwater flux from ice sheets is applied and quantified. Given this context, I feel it would be beneficial to emphasize the importance of using improved meltwater forcing to conduct Holocene experiments in the latest PMIP-CMIP models.

Reply: We have added the following sentences to the Conclusions to address this point:

(L416-419) However, confidence in this interpretation remains limited by uncertainties in the spatial and temporal distribution of the applied meltwater flux, and by biases in the deglaciation simulations themselves. The difficulty of discerning the impact of meltwater during the MH therefore reflects the uncertainties of current climate models and experimental designs (Sect. 4.2), rather than constituting evidence against a role for meltwater.

(L423-424) Further investigation will require Holocene experiments with improved meltwater forcing in the latest generation of PMIP-CMIP models.

L393: This statement seems to contradict what is written in L159–160. Could the authors please clarify, or have I misunderstood the text?

Reply: We agree the sentence in L159-160 was a bit misleading. When we stated that the PMIP3 ensemble seems to have “less outliers” than the PMIP4 ensemble, we were only referring to the seasonal cycle of MLD in three specific deep-water formation regions (Fig. 3 and 4). Here at L393, we are referring to the winter MLD anomaly maps (Fig. 5), which show substantial improvement in the PMIP4 ensemble, with “reduced inter-model spread” as can be quantified by the standard deviation and “more realistic locations of deep convection”, than in the PMIP3 ensemble. Therefore, we have removed the sentence at L159-160, and added a citation to Fig. 5 at L393 (now L408) to avoid confusion.

Fig. 4: Please use the same scale for the y-axis in panels (b) and (e).

Reply: Thank you for pointing this out, we have made the change.

Report #2

Referee: Marlene Klockmann

1. It seems that a reference for the CMIP5/PMIP3 protocol is missing. The references in line 76 only refer to CMIP6/PMIP4. I am not sure which is the best reference, but Taylor et al 2012 (https://doi.org/10.1175/BAMS-D-11-00094.1) could be a good starting point.
Reply: Thank you for noticing this omission. Taylor et al. (2012) describe the CMIP5 experimental design as a whole; for the PMIP3 mid-Holocene protocol specifically, we cite Braconnot et al. (2011), which we located through the Taylor et al. (2012) paper. Both references are added at L77.

2. It is unfortunate that the transient LDv1 experiments do not show any improvements over the equilibrium simulations and thus cannot directly support the hypothesis that adding meltwater flux would improve the MLD response. This is reflected in the discussion, but I don't find the corresponding sentence in the conclusion too vague (l.399: "Comparison against simulations of the last deglaciation may support this hypothesis, although large uncertainties and biases reduce the confidence in this interpretation.") Perhaps you can revise this sentence to state more clearly what of the LDv1 experiments supports the hypothesis (e.g, weaker overestimation of Labrador Sea MLDs?) and which uncertainties reduce the confidence (e.g. meltwater distribution?).

Reply: We have revised this part of the Conclusions as suggested (L414-419):

Comparison against simulations of the last deglaciation lends some support to this hypothesis, in that the overestimation of Labrador Sea MLD is weaker in these experiments (Figs. 3, 4, 6). However, confidence in this interpretation remains limited by uncertainties in the spatial and temporal distribution of the applied meltwater flux, and by biases in the deglaciation simulations themselves. The difficulty of discerning the impact of meltwater during the MH therefore reflects the uncertainties of current climate models and experimental designs (Sect. 4.2), rather than constituting evidence against a role for meltwater.