We thank both reviewers for their constructive and helpful comments, which helped strengthen the manuscript. We will improve the text and the figures as suggested by the reviewers. We will also conduct additional analyses of the PMIP3 MH and PMIP4 deglaciation experiments to provide a more comprehensive assessment of proxy-model agreement.

The original comments are displayed in blue, and our replies in black.

RC1 Marlene Klockmann

Main

Intro/literature context: The Introduction is very short and remains quite superficial. In particular, the context of previous literature is missing. I understand that this is the first model-proxy comparison for MH MLDs, and also the first PMIP4 MLD model intercomparison. However, previous studies have looked at changes in deep convection or MLDs in individual models or proxies (e.g. Thornally et al, 2013, https://doi.org/10.5194/cp-9-2073-2013; Otto-Bliesner et al, 2020, https://doi.org/10.1029/2020PA003957). What have they found and what open questions did they leave? Also, the MLD reconstruction has already been published, so a short summary of your previous study could also be helpful in the introduction to set the scene.

We will improve the introduction as suggested by the referee. We plan to add the following text to the introduction:

Past changes in deep convection have been difficult to reconstruct, as the existing proxies are not direct indicators of convection. Recently, Wu et al. (2025) suggested that dinoflagellate cyst (dinocyst) assemblages can be influenced by wintertime deepening of the MLD in response to deep convection events, allowing for the quantitative reconstruction of MLD from dinocyst records recovered from marine sediment cores. The results from the study by Wu et al. (2025) indicate a westward migration of deep convection centers in the subpolar North Atlantic around 6 ka BP, and a late onset of modern-like deep convection in the Labrador Sea about 4 ka BP. This is compatible with previous studies using bottom water proxies such as sortable silt and benthic foraminifera to reconstruct the overflow of deep water from the Nordic Seas via the Iceland-Scotland Ridge, which indicate that deep-water formation rate in the Nordic Seas reached a maximum at 6-7 ka BP before decreasing progressively (e.g., Kissel et al., 2013; Rasmussen et al., 2002; Thornalley et al., 2013). Other studies, such as those of Renssen et al. (2005) and Thornalley et al. (2013), used transient simulations with individual models forced by Holocene changes in orbital parameters, Greenhouse gas concentrations, and ice sheet meltwater to simulate changes in winter convection depth in the Nordic Seas and the

Labrador Sea. These studies also suggest a decrease in deep convection in the Nordic Seas and an opposite trend in the Labrador Sea during the Holocene. However, the reconstructions of Wu et al. (2025) indicate an abrupt change of winter MLD in these areas rather than a long-term gradual shift as shown in the simulations; additionally, the signal of deep convection in the Labrador Sea appeared much later in the reconstructed MLD than in the simulations.

Calendar adjustment: Because of the different orbital configuration, the seasons had different lengths during the MH than PI. If monthly or seasonal variables are considered, I understand that a calendar adjustment should be performed to account for that (Bartlein & Shafer, 2019, https://doi.org/10.5194/gmd-12-3889-2019). This was also done, e.g. in Brierly et al 2020. I would expect this to be relevant also for seasonal variables like sea-ice and MLD. Did you also perform a calendar adjustment? And if not, could you shortly justify why this is not necessary here?

This is a relevant point that we didn't consider. However, after careful examination of the relevant section in Brierley et al. (2020), we believe that a calendar adjustment might not be necessary for MLD. Calendar adjustments could affect the accuracy of variables that change abruptly throughout the year rather than gradually, such as monsoon variables. Brierley et al. (2020) conducted a test using daily and monthly monsoon data, calculating the "true" seasonal average on the paleo-calendar using daily data. The results showed that the monthly monsoon data without calendar adjustment is more accurate than calendar adjusted. Unfortunately, no daily-resolution salinity and temperature data, or online-calculated MLD data, are available from PMIP4, so we cannot apply the same approach for validation. Nevertheless, given the similarity in temporal distribution between monsoon precipitation and MLD (seasonal high values and abrupt changes rather than gradual throughout the year), we believe that calendar adjustment would likely introduce interpolation errors into the MLD data as well. Sime et al. (2025) also stated that calendar adjustment would lead to incorrect lengths of the melt, ice-free, or ice-growth seasons.

Focus on North Atlantic only: I recommend to exclude the Southern Ocean, both from the text and the figures. The main focus of the study is the North Atlantic, so the short passages on the Southern Ocean are only disrupting the flow of the story. Excluding the SO also from the figures will greatly improve readability of the figures (see comments on Fig 1, S2 and S3).

We will exclude the Southern Ocean from the text and the figures as suggested.

Sea-ice reconstruction: There are no details given for the sea-ice reconstruction. What is the reconstruction based on? It seems as if it was based on the same cores, but on which

proxies. Did you perform it yourself or was it previously published? Have you performed the similarity test also for sea ice? And would the model with the highest similarity score also have the best sea-ice agreement?

We performed the sea ice reconstruction using the same proxy (dinocyst assemblage data) and the same method (modern analogue technique with the R package 'analogue') as for the MLD reconstruction. We will add these details to the methods section. Regarding the referee's question "would the model with the highest similarity score also have the best sea-ice agreement?", we will run a Cohen's kappa test to explore this.

Discussion of uncertainties: Please add some discussion of the proxy-related uncertainties. What are the proxy-related uncertainties? How confident are you in both the MLD and the sea-ice reconstruction? In the discussion you only state "The model biases are challenging to analyse here as there is uncertainty in the reconstructed MLD data as well (see Wu et al., 2025)." This is not sufficient. Did you take the uncertainty into account? Can you quantify it? Is it comparable to the uncertainty of the multi-model mean (MMM)?

The main sources of proxy-related uncertainties are errors in sediment sample age estimation and errors in the reconstruction method. The errors in sediment core age modelling vary from core to core and depend on the number of ^{14}C dating points in each core and the accuracy of the estimated ocean reservoir-age correction. This error can be as high as a few hundred years, affecting the actual 5.5-6.5 ka BP time window used to represent the MH in sediment cores. Its impact on the results, however, is difficult to analyze and core-specific. The error of reconstruction, root mean square error of prediction (RMSEP), is estimated from leave-one-out cross-validation to be 40.9 m for the winter MLD, and 17.6% for the winter sea-ice concentration. Finally, we found that the reconstruction method can exhibit regional biases, as discussed in Wu et al. (2025), of which we hope to reduce the effect by using the MH minus PI anomaly approach. For example, our reconstruction method tends to underestimate the MLD in the southern Labrador Sea and the Iceland Basin. Overall, the proxy-related uncertainties are probably comparable to, or slightly less than, the uncertainty of the multi-model mean.

The above text will be added to the discussion.

Minor

We will adjust the manuscript and figures following the referee's suggestions. Here is our reply to some specific points:

l.53-59: Also from reading Wu et al 2025, I understood that WOA18 is the best choice for the calibration? Why switch to Boyer-Montegut and are the results very sensitive to the calibration data set?

Yes, WOA18 yielded the best results in cross-validation tests, but the difference in performance between different MLD datasets is rather small. The calibration dataset would have a direct impact on the reconstructions. The reconstructed value is calculated as the distance-weighted mean of the 5 closest modern analogues. As different calibration dataset assigns different MLD values to the modern analogues, the reconstructed value would change correspondingly. The fundamental difference between different MLD datasets is spatial coverage and the definition of MLD. Here, the choice of de Boyer-Montegut (2023) (hereafter BM23) is a balance between accurate estimation of MLD from model outputs and performance of the reconstruction method. BM23 has a full coverage of the global oceans which ensures accurate assignment of MLD values to the calibration dinocyst data points. The MLD criterion used in BM23 is recommended by Treguier et al. (2023) for model intercomparison as a cost-efficient and relatively accurate estimation of the MLD. Hence, using BM23 for reconstruction allows us to use a consistent MLD definition across proxy reconstruction and model simulation which we find to be optimal. We will add these details to the text.

Fig.2: I find the line for the "modern data" in the MH panels somewhat misleading. Consider removing it.

We will remove the line for modern data in the MH panels and replace it with reconstructed maximum MLD values for the 5.5-6.5 ka BP interval from nearby proxy records (except for the Irminger Sea, where we do not have records nearby). As suggested by Referee 2, the reconstructed maximums might correspond to periods of lower meltwater influence and therefore be more comparable to the MH model simulations. We believe that adding the reconstructed maximums to the MH panels could offer a better overview of how reliable/biased the models are.

l.130-131 & l.139-148: How do these two parts go together? It is very difficult to see from Fig.1 and S2, whether the general statement in l.130-131 holds (see comments on suggestions for Fig.1, S2 and S3). From comparing Fig.3a and Fig.S4, the general statement does also not really hold everywhere for the MMM. And, the individual model descriptions in l.139-148 also seem to show that the sea-ice extent and the MLD are unrelated in many models.

The figures will be reworked to focus on the North Atlantic area. In lines 130-131, we stated that a negative correlation between anomalies in winter sea ice concentration and MLD is typically seen; and in lines 139-148, we discussed that there can be exceptions because the initial state of sea ice concentration in PI can affect how its changes (anomalies) relate to the MLD. In areas already densely covered by sea ice in PI, even if there is an increase in sea ice in MH, the MLD changes very little since deep convection was already inhibited in

PI, so the MLD remains at a value close to zero in MH. On the other hand, in ice-free areas of PI, deep convection becomes stronger with MH forcing due to enhanced ocean heat loss under reduced winter insolation, even though sea ice remains unchanged. And in these areas (ice-free in PI), increased MLD along with no change in sea ice is often the case, as intense deep convection decreases stratification and hence conditions remain unfavorable for local sea ice formation even though winter is colder in MH. Despite these exceptions related to the initial states of the sea ice cover, the two-way feedback between sea ice and open-ocean convection we discussed at lines 130-131 remains valid.

l.180: Fig.S4 only shows anomalies not the absolute values.

Thanks for pointing that out. The wrong figure is cited here; we will attach the correct figures to the supplementary materials.

l.215-217: Would higher-frequency output be beneficial for the model-proxy comparison? Can the proxies resolve extreme events?

Because the MLD is a nonlinear function of the density profile, computing it from lower-frequency output could introduce uncertainty and might fail to represent the true simulated MLD. While the proxies do not resolve individual extreme events, these events could leave a footprint or diluted signal in the sediment layers that is included in each sample. It is unclear whether higher-frequency output would be beneficial for the proxy-model comparison. However, the usage of lower-frequency output probably contributed to the model uncertainty.

l.123: Not sure, that Fig.2 supports this statement of a "realistic" seasonal cycle. All models have a seasonal cycle in MLD, but most models either strongly under- or overestimate the seasonal cycle. Is that realistic?

We find the simulated seasonal cycles “realistic” in the sense that MLDs slowly deepen starting from late autumn and rapidly shoal in spring as seen in modern observations, despite the models’ under- or overestimation of the maximum strength of wintertime convection. We will add this to the manuscript.

RC2 Sam Sherriff-Tadano

Main Comments

1. Inclusion of PMIP3 and deglaciation simulations

As this appears to be the first study conducting a proxy–model comparison of MLD for the mid-Holocene, it would be highly beneficial to include simulations from PMIP3 and

available deglaciation runs in the analysis. Incorporating these outputs would allow for a more comprehensive assessment of model robustness and inter-model spread. Furthermore, including deglaciation simulations would provide the necessary data to validate the authors' hypothesis (L182) that freshwater forcing is a key driver of MLD shoaling in the Labrador Sea.

We will add the PMIP3 model outputs to the analysis. After a preliminary investigation, we found 9 models that have the required output for MLD calculation: MPI-ESM-P, MRI-CGCM3, BCC-CSM1.1, CNRM-CM5, IPSL-CM5A-LR, FGOALS-g2, FGOALS-s2, GISS-E2-R, MIROC-ESM. For the last deglaciation simulations, we will aim for a proxy-model comparison of Holocene time series (12-0 ka). Therefore, simulations must have an adequate temporal coverage of this time interval. There are a few deglaciation runs that meet this requirement: TraCE-21ka, the HadCM3B runs, and the MPI-ESM-CR runs. However, data availability is often more limited than in the midHolocene experiment and does not allow recomputation of the MLD. Hence, for the deglaciation experiments, we will have to use the models' online-calculated MLDs, even though their definitions differ.

2. Discussion of proxy and modeling uncertainties

I am concerned about the sensitivity of the results presented in L112–121 to uncertainties in the proxy data and the specific definition of MLD used. Based on the current analysis, it appears that models with the smallest MLD changes yield smaller RMSE values (e.g., Fig. 1). If this is the case, how should the RMSE be interpreted? The authors should carefully reconcile this by discussing the impacts of proxy uncertainty and MLD definitions on their findings.

In this regard, please provide the uncertainty ranges for the reconstructed MLD for both the Mid-Holocene (MH) and Pre-Industrial (PI) periods (L56–58). One approach would be to show the maximum/minimum MLD over the 5,500–6,500 BP time domain. Since the PMIP4 models lack freshwater forcing, it may be more reasonable to compare the models against the maximum MLD of the proxy records, as these likely represent periods with minimal impact from ice discharge.

The reviewer is right that models with the smallest MLD changes yield smaller RMSE values. This is especially true in the Labrador Sea, where most models either simulate no change in MLD or a positive change, which is the opposite of what proxy records indicate. As models often simulate MLD changes in the wrong direction, models that simulate smaller changes become advantageous in the RMSE analysis. Therefore, the RMSE is probably not a good indicator of proxy-model agreement here and should only be interpreted as a supplementary measurement.

The proxy reconstruction error, or root mean square error of prediction (RMSEP), is estimated from cross-validation to be 40.9 m for the winter MLD, which is smaller than all proxy-model RMSE values.

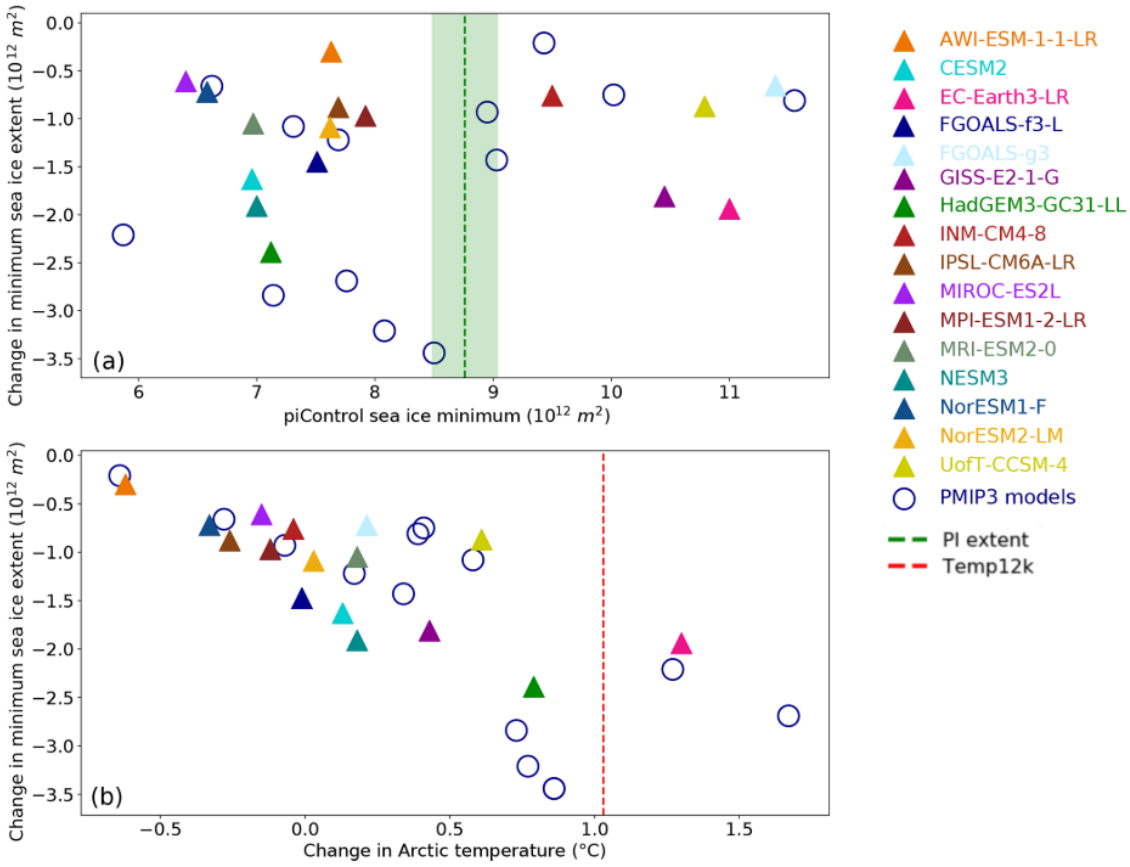
The above details and further discussion will be added to the manuscript. We will also explore the idea of showing maximum MLD over the MH time slice from proxy reconstructions. For now, we plan to add the reconstructed maximums to the MH panels in Figure 2, to replace the "modern observation" curve.

Specific Comments

We will address the points raised by the reviewer as suggested. Here is our reply to some specific questions:

L82–83: Could the model spread be related to the magnitude of Arctic warming? For instance, do IPSL and NESM3 exhibit weaker Arctic warming compared to the others?

Looking at the evaluation of the PMIP4 midHolocene experiment by Brierley et al. (2020), IPSL and NESM3 do not stand out among the models in terms of Arctic warming (see figure from Brierley et al., 2020 below). In the MH, IPSL simulates a slightly lower Arctic annual mean temperature, whereas NESM3 simulates a slightly higher one.



L155–161: Does this imply that the vegetation feedback enhances the AMOC, thereby causing the retreat of Arctic/Labrador sea ice? The spatial map of sea-ice anomalies in Fig. S2 seems consistent with this view (showing reduced sea ice in the North Atlantic and increased sea ice in the Southern Ocean).

We believe that the causal chain goes as follows: vegetation feedback alters the moisture balance, making the North Atlantic surface water saltier, which decreases water-column stability, thereby enhancing deep convection, as evidenced by deeper MLDs in the Labrador Sea and Irminger Sea. The stronger deep convection and the associated deep-water formation then lead to a more intense AMOC and sea ice retreat, further strengthening the feedback loop.