We thank both reviewers for their constructive and helpful comments, which helped strengthen the manuscript. We have improved the text and the figures as suggested by the reviewers. We have also conducted 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. The line numbers mentioned in our replies refer to the marked-up version of the revised manuscript.

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 have added the following text to summarise previous studies (L55-67):

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 centres 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.

Since we add the PMIP3 MH and PMIP4 last deglaciation to the analysis, we have also added background information on the difference between PMIP3 and PMIP4 (L35-39):

Compared to the previous PMIP3 *midHolocene* experiments, which only imposed the change in orbital configuration, the PMIP4 protocol includes a realistic specification of atmospheric greenhouse gas concentrations (Otto-Bliesner et al., 2017). Additionally, the PMIP4 simulations represent a new generation of climate models with higher complexity (e.g., dynamic vegetation, ocean biogeochemistry), improved parameterisations and resolution (Brierley et al., 2020; Xu et al., 2025).

And the potential influence of meltwater forcing (L46-54):

While PMIP equilibrium simulations provide a benchmark for mean-state changes, they may not accurately capture important feedbacks due to uncertainties in the boundary and initial conditions (Otto-Bliesner et al., 2017). The known uncertainties include, for instance, the prescription of modern vegetation cover and ice sheets, in addition to the lack of transient history of freshwater release from ice sheets into the ocean. Marine proxy records have demonstrated a significant impact of ice-sheet meltwater on ocean circulation during the last deglaciation (e.g., Alley and Clark, 1999; and references therein). Various model-based studies of the modern ocean also show that increasing meltwater from Greenland can lead to significant slowing of the AMOC (e.g., Golledge et al., 2019), and has the potential of disrupted deep-water formation over extended time scales (Hoshyar et al., 2025). In particular, Pontes and Menviel (2024) show that improved data-model agreement on AMOC variations over the historical period is achieved when an upper-end meltwater input is included in the simulations.

Finally we have added the following text to complete the objectives of this study (L71-73):

We also include 13 models from the previous PMIP3 ensemble to allow for a comprehensive assessment of model robustness across different model generations. We supplement our analysis with transient simulations of the last deglaciation (PMIP4 LDv1; Ivanovic et al., 2016) to assess the impact of ice-sheet meltwater feedback.

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.

We have added this sentence to clarify that calendar adjustment was not performed (L86-87):

No calendar adjustments were applied because they could affect the accuracy of variables that change abruptly throughout the year rather than gradually (Brierley et al., 2020).

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 have excluded 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 have added this detail to the methods section (L105): "The MLD and sea ice reconstructions are based on …".

Regarding the referee's question "would the model with the highest similarity score also have the best sea-ice agreement?", we have run a Cohen's kappa test to explore this. The results are presented in Table S1 and discussed in the text (L292-296):

Uncertainties in the PI sea ice extent are likely a more important source of the inter-model spread in MLD than uncertainties in the response of sea ice concentration. Our statistical analyses on winter sea ice concentration (Table S1) show no universal correlation between model performance in simulating MLD anomaly and sea ice anomaly: significant agreement on sea ice concentration anomaly does not guarantee agreement on MLD anomaly (e.g., FGOALS-g3, MPI-ESM1.2-LR), although the model with the best performance for MLD anomaly (MRI-ESM2.0) does have a significant agreement for sea ice anomaly. This highlights the importance of other mechanisms affecting simulated MLD.

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)?

We have added the following text to the discussion to address the proxy-related uncertainties (L391-400):

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 6.5-5.5 ka BP time window used to represent the MH in sediment cores. Its impact on the results, however, is difficult to analyse and core-specific. The reconstruction error (RMSEP) is estimated from leave-one-out cross-validation to be 40.9 m for winter MLD, 16.0 m for annual MLD, and 17.6% for winter sea ice concentration. Finally, we found that the reconstruction method can exhibit regional biases, as discussed in Wu et al. (2025), which we attempt to mitigate using the MH minus PI anomaly approach. For example, the reconstruction method tends to underestimate the MLD in the southern Labrador Sea and the Iceland Basin (see Fig. S12 in Wu et al., 2025). Overall, the proxy-related uncertainties are probably comparable to, or slightly less than, the uncertainty of the multi-model mean.

Minor

l.36-37: Shortly elaborate which parts of the global climate are simulated well.

We have added a brief summary to address this comment (L42-44):

In response to these external forcings, models simulate some global climate changes that are broadly consistent with theory and reconstructions, such as warmer Northern Hemisphere summer and associated shifts in tropical rainfall, reduced Arctic summer sea ice extent, cooler annual mean temperature globally but warmer in the Arctic, etc. (Brierley et al., 2020).

l.53-59: Please shortly explain what the modern analogue technique does. 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.

A short explanation of the modern analogue technique is added (L108-110):

The modern analogue technique is based on the similarity between fossil assemblages and modern assemblages, using the closest modern analogues to a fossil sample to predict the corresponding environmental parameters of the past.

Fig.1: Limit the extent of the map to the North Atl., perhaps replace each global map with a smaller map as in Fig.3. Also include the proxies in each of the maps? Also please use a discrete colour-bar similar to Fig.3.

We have updated the figure as suggested.

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

We have removed the line for modern data in the MH panels and replaced it with reconstructed maximum MLD values over the 5.5-6.5 ka BP interval from nearby proxy records (except for the Irminger Sea, where we do not have records in the area). 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 have added the reconstructed maximums to the MH panels to offer a better overview of how reliable/biased the models are.

l.109-111: "recurring pattern", please be specific and say which pattern is captured by the MMM.

We have improved the sentence as suggested (L195-196): "... the recurring pattern of increase in the Labrador Sea and northwestward shift in the Nordic Seas found in individual models ...".

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 have been 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.

We have adjusted the second part to be clearer (L278-286):

The initial state of sea ice concentration in PI can affect how its changes relate to the MLD. Unlike the previously mentioned negative correlation between MLD and sea ice concentration, when an increase in sea ice takes place in already densely sea ice-covered areas, there is no evident change in MLD since deep convection was already inhibited in PI and MLD remains very shallow (e.g., EC-Earth3, EC-Earth3-LR, INM-CM4.8, MIROC-ES2L). In contrast, in an ice-free area, there is generally an increase in convection due to enhanced ocean heat loss under reduced winter insolation, even though there is no change in sea ice (e.g., CESM2, GISS-E2.1G, MPI-ESM1.2-LR, MRI-ESM2.0, NorESM2-LM). In the areas that are ice-free in PI, an increased MLD without change in sea ice is often seen, as intense deep convection decreases stratification and hence conditions remain unfavourable for local sea ice formation.

l.175-181: How confident are you in the sea-ice reconstructions (see also main comments on sea-ice reconstruction and proxy-related uncertainties)?

We have added the information on the error of the reconstructions in the methods section (L113-115):

Based on cross-validation tests using the leave-one-out method, the root mean square error of prediction (RMSEP) for the reconstructions is estimated to be 40.9 m for winter MLD, 16.0 m for annual MLD, and 17.6% for winter sea ice concentration.

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

Thanks for pointing that out. The wrong figure is cited here; we have added the correct figure to the supplementary materials (the name is still Fig. S4 because the previous Fig. S4 has been moved to the main text).

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 have added this to the manuscript (L157 and L406).

Fig.S2 and S3: same comment as to Fig.1. Limit the map to the North Atlantic. Consider adding the proxies also to Fig.2, and again use a discrete colour-bar. And please reverse the colour-bar to have red indicating a sea-ice increase. I think I understand why red was chosen for a sea-ice decrease, but I automatically read "red" as a positive anomaly and had to re-think multiple times while looking at Figures S2 and S4.

We have made this change.

I would suggest to include Fig S4 in the main manuscript. Please reverse the colour-bar also here to have red indicating a sea-ice increase.

We have made this change. This figure has been added to the new Figure 5.

Throughout the text, "observed" or "observation" is often used for a statement that is based on simulated results. I suggest to change the wording, as "observations" are immediately associated with actual measurements on the actual world.

We have made this change.

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 have added the PMIP3 model outputs to the analysis. We have found 13 models that have the required output for MLD calculation (presented in Table 2). For the last deglaciation simulations, we have carried out a proxy-model comparison of Holocene time series (12-0 ka) and the MH time slice. 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 (presented in Table 3). 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 had to use the models' online-calculated MLDs, even though their definitions differ.

In the introduction, we have added background information on the difference between PMIP3 and PMIP4 (L35-39):

Compared to the previous PMIP3 *midHolocene* experiments, which only imposed the change in orbital configuration, the PMIP4 protocol includes a realistic specification of atmospheric greenhouse gas concentrations (Otto-Bliesner et al., 2017). Additionally, the PMIP4 simulations represent a new generation of climate models with higher complexity (e.g., dynamic vegetation, ocean biogeochemistry), improved parameterisations and resolution (Brierley et al., 2020; Xu et al., 2025).

And the potential influence of meltwater forcing (L46-54):

While PMIP equilibrium simulations provide a benchmark for mean-state changes, they may not accurately capture important feedbacks due to uncertainties in the boundary and initial conditions (Otto-Bliesner et al., 2017). The known uncertainties include, for instance, the prescription of modern vegetation cover and ice sheets, in addition to the lack of transient history of freshwater release from ice sheets into the ocean. Marine proxy records have demonstrated a significant impact of ice-sheet meltwater on ocean circulation during the last deglaciation (e.g., Alley and Clark, 1999; and references therein). Various model-based studies of the modern ocean also show that increasing meltwater from Greenland can lead to significant slowing of the AMOC (e.g., Golledge et al., 2019), and has the potential of disrupted deep-water formation over extended time scales (Hoshyar et al., 2025). In particular, Pontes and Menviel (2024) show that improved data-model agreement on AMOC variations over the historical period is achieved when an upper-end meltwater input is included in the simulations.

And to complete the objectives of this study (L71-73):

We also include 13 models from the previous PMIP3 ensemble to allow for a comprehensive assessment of model robustness across different model generations. We supplement our analysis with transient simulations of the last deglaciation (PMIP4 LDv1; Ivanovic et al., 2016) to assess the impact of ice-sheet meltwater feedback.

In the methods section, we have added the following descriptions (L88-104):

We also included the PMIP4 last deglaciation experiment (LDv1) for process-based validation related to transient meltwater forcing. The experimental design and boundary conditions are described in Ivanovic et al. (2016). A full analysis of the participating models is presented in Snoll et al. (2024a). We only analysed simulations covering the MH in this study (Table 3). The simulations are prescribed with time-varying orbital forcing, greenhouse gases, ice sheets and other geographical changes. The GLAC-1D (Ivanovic et al., 2016) and ICE-6G_C (Peltier et al., 2015) are the two recommended ice-sheet reconstructions in the PMIP4 last deglaciation protocol, although the older ICE-5G (Peltier, 2004) is also used by some models. The protocol recommends calculating the meltwater flux based on ice volume change from the chosen ice-sheet reconstruction and applying this flux either as a globally uniform freshwater flux (*melt-uniform*) or at individual meltwater river outlets along the coast (*melt-routed*). However, the meltwater input used in the TraCE-21ka simulation does not follow the ice-sheet reconstructions mentioned above; in this simulation, a freshwater flux is used to regulate the AMOC during the early last deglaciation, so that the model produces climate changes as recorded by proxies (Liu et al., 2009; Snoll et al., 2024a). The three versions of the MPI-ESM-CR model, P1, P2 and P3, correspond to different model physics including tuning of parameters, parameterisations, and bug fixes (Kapsch et al., 2022). The output available from these long transient simulations is sometimes limited to annual and decadal averages. Recomputing the MLD from annual or decadal temperature and salinity data would probably introduce more bias than benefits since MLD is a nonlinear function of the density profile (e.g., de Boyer Montégut et al., 2004; Treguier et al., 2023). Therefore, we used the LDv1 models' online-calculated MLDs, even though their definitions differ.

And (L120-121):

For the transient deglaciation simulations, we use the same 6.5-5.5 ka BP window to represent the MH as for the proxy records, and the average of 0.1-0 ka BP as the PI baseline value.

The results of the PMIP3 ensemble are presented throughout the Results section and in Figures 2 and 4; those of the last deglaciation simulations are presented in the new section "3.3 Transient simulations of the last deglaciation" (L229-244) and in Figures 6 and 7. Discussion on the PMIP3 and the last deglaciation simulations has been added at L256-263, L360-373, L383-387.

The main conclusions related to the PMIP3 and deglaciation simulations have been added as well (L407-408):

Compared to the PMIP3 ensemble, the PMIP4 ensemble shows substantial improvement with reduced inter-model spread and more realistic locations of deep convection in the North Atlantic.

At L410-412:

While there is relatively good agreement in the Nordic Seas, significant discrepancies in the Labrador Sea persist from PMIP3 to PMIP4. This confirms that the discrepancies in the Labrador Sea are not artefacts of a single model generation, but rather a systematic bias in the MH simulations.

And at L414-416:

Comparison against simulations of the last deglaciation may support this hypothesis, although large uncertainties and biases reduce the confidence in this interpretation.

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 have been added to the manuscript (L199-203).

We have added the reconstructed maximums to the MH panels in Figures 2 and 4, to offer a better overview of how reliable/biased the models are. At most proxy sites, using maximum instead of mean does not significantly change the MLD anomaly. This is because many proxy records have a low temporal resolution with only a few data points within the MH interval.

We have added the following text to the discussion to address the proxy-related uncertainties (L391-400):

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 6.5-5.5 ka BP time window used to represent the MH in sediment cores. Its impact on the results, however, is difficult to analyse and core-specific. The reconstruction error (RMSEP) is estimated from leave-one-out cross-validation to be 40.9 m for winter MLD, 16.0 m for annual MLD, and 17.6% for winter sea ice concentration. Finally, we found that the reconstruction method can exhibit regional biases, as discussed in Wu et al. (2025), which we attempt to mitigate using the MH minus PI anomaly approach. For example, the reconstruction method tends to underestimate the MLD in the southern Labrador Sea and the Iceland Basin (see Fig. S12 in Wu et al., 2025). Overall, the proxy-related uncertainties are probably comparable to, or slightly less than, the uncertainty of the multi-model mean.

Specific Comments

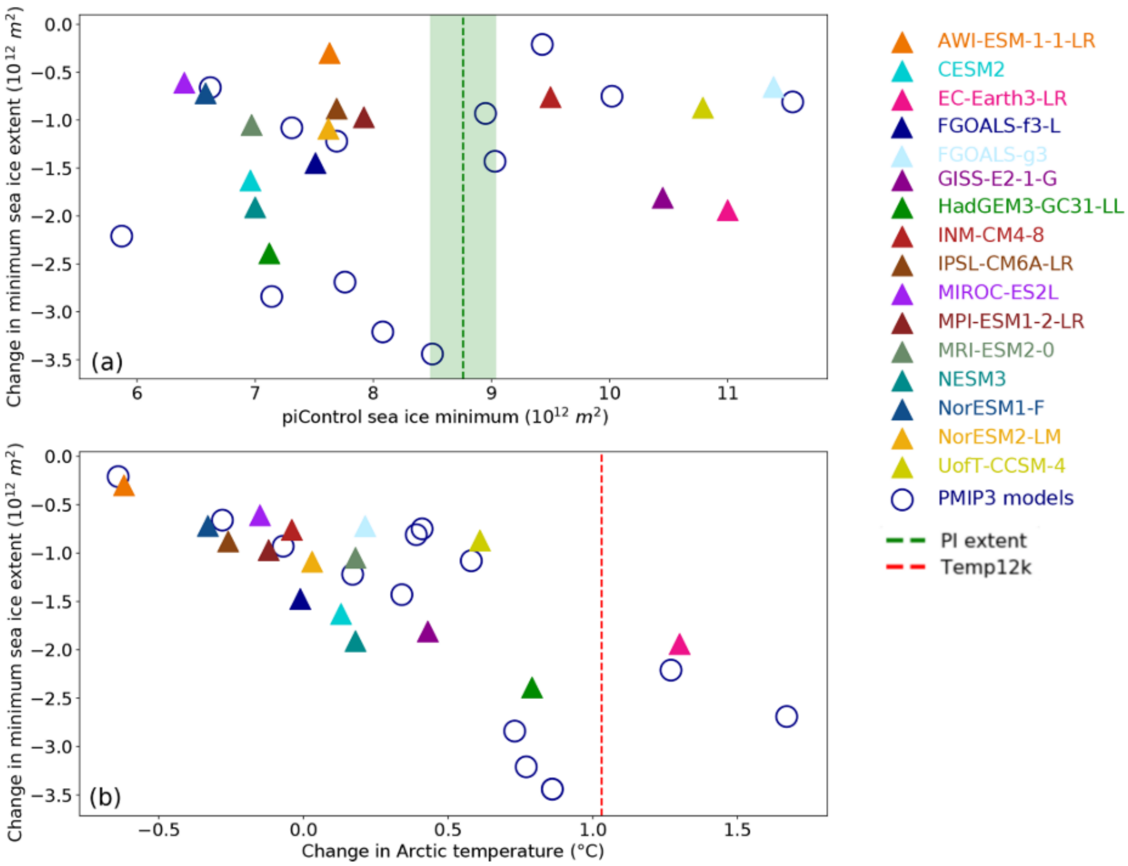
L36: Could you elaborate further on this sentence (e.g., providing both global and Arctic perspectives)?

We have added a brief summary to address this comment (L42-44):

In response to these external forcings, models simulate some global climate changes that are broadly consistent with theory and reconstructions, such as warmer Northern Hemisphere summer and associated shifts in tropical rainfall, reduced Arctic summer sea ice extent, cooler annual mean temperature globally but warmer in the Arctic, etc. (Brierley et al., 2020).

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). Additionally, vegetation feedback can affect Arctic sea ice by modifying planetary albedo (e.g., O'ishi and Abe-Ouchi 2011 https://doi.org/10.1029/2011GL048001; O'ishi et al. 2021 https://doi.org/10.5194/cp-17-21-2021); these mechanisms should be explicitly mentioned.

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.

We have added the following text to address this comment (L305-309):

In the EC-Earth simulations, this high salinity anomaly appears to affect the Labrador Sea the most, reducing stratification and sea ice formation, promoting convection and deep-water formation, and possibly enhancing the AMOC, which further strengthens the feedback loop. Vegetation can also increase warming in the Arctic through albedo feedback, leading to amplified Arctic sea ice retreat and reduced sea ice export in winter (e.g., O'ishi and Abe-Ouchi, 2011), which reinforce the stronger wintertime convection in the Labrador Sea.

L162–163 and L172: Please clarify these points. Do the authors mean that the differences between EC-Earth3 and IPSL-CM6A-LR reflect opposing AMOC responses?

Yes, according to Jiang et al. (2023), EC-Earth3-LR simulates a stronger AMOC in the MH relative to PI while IPSL-CM6A-LR simulates a slightly reduced one, which is consistent with the fact that EC-Earth3-LR has stronger positive anomalies of MLD than negative anomalies, while IPSL-CM6A-LR has more negative anomalies.

We have added this sentence to clarify (L324-325): For example, EC-Earth3-LR simulates a stronger AMOC in the MH relative to PI while IPSL-CM6A-LR simulates a slightly reduced one (Jiang et al., 2023).

Fig. 1: I agree with Referee 1 regarding the removal of MLD maps over the Southern Ocean to keep the focus on the North Atlantic/Arctic.

We have made this change.