Thank you for the manuscript "The North Atlantic mean state in eddy-resolving coupled models: a multimodel study". Understanding how increasing resolution improves the mean state is an important topic and I welcome this study.

We would like to sincerely acknowledge the reviewer for their constructive comments that have greatly contributed to improve this manuscript. Please, find below the answers to all the points raised in the comments (in blue). Some of the changes in the manuscript have already been included in green.

However, I have some major comments so have to recommend major revisions. There are a lot of statements throughout about quantities being larger/smaller in the HR ensemble compared to the LR ensemble, however this isn't really tested. It seems to be based on differences in the ensemble mean without taking into account the fact that the two ensembles are quite different in size and possibly quality. If 4 members were picked at random from the LR ensemble, what is the probability that they look like the HR ensemble? It should also be borne in mind that the CMIP6 ensemble includes some models which are a long way from the observations. The authors should statistically test their assertions – one way of doing this is by using single number metrics (for instance the AMOC strength, SST in the CNA etc) and testing whether the ensembles are statistically different by: randomly picking 4 members from the LR ensemble, calculating the ensemble mean, then repeating until you have a distribution from the LR ensemble. This will show whether the HR ensemble is really different from the LR ensemble, or whether there are some LR members which have similar properties.

To address this point, we have applied "bootstrapping" to different single number metrics associated with the variables analyzed in the manuscript: SST and SSS biases in the LS, CNA, and NCH regions; max. AMOC, as well as AMOC RMSE and AMOC correlation to RAPID at 26.5ºN; RMSE and correlation of temperature, salinity and density profiles in the LIS box compared to EN4; max. MLD in the LIS box; and max. strength of the SPG (first column in Table T1).

We have allowed repetition (replacement) in the bootstrapping samples from both the LR and HR ensembles, to better describe the variability of the ensembles. Significance has been assessed by calculating the 95% confidence intervals (CIs) of the distribution of the differences in means between the two ensembles: mean(HR) - mean(LR).

First, bootstrapping has been applied using maximum ensemble sizes in both the LR and the HR ensembles (second column in Table T1). Here is the pseudocode:

```
size_LR = total size of the LR ensemble

size_HR = total size of the HR ensemble (it is usually 4)

distribution = {}
for i from 0 to 9999:
    sample_LR = random sample from the LR ensemble, of size size_LR, taken with replacement
    sample_HR = random sample from the HR ensemble, of size size_HR, taken with replacement
    difference = mean(sample_HR) - mean(sample_LR)
    add difference to distribution
confidence_interval = [2.5th percentile of distribution, 97.5th percentile of distribution]
```

Subsequently, another analysis has been performed after reducing the size of the LR ensemble samples to the size of the HR ensemble, i.e. by assigning *size_LR* = total size of the HR ensemble (third column in Table T1).

If the CI obtained in this second analysis did not contain 0, we repeated the analysis by gradually increasing the size of the LR ensemble samples until a CI falling entirely to the right (or to the left) of zero was obtained (last column in Table T1).

For the LS SST and SSS biases, when bootstrapping is applied employing LR samples with the total LR ensemble size, the difference in means between the two ensembles is significant (i.e. the CI obtained does not include 0; Table T1). By contrast, when the size of the LR ensemble samples is reduced to the size of the HR ensemble (i.e., to 4), the CI does include 0. Sizes of 19 and 25 for the LR ensemble samples are required for SST and SSS, respectively, for the difference in means to become significant. This happens because there are several models in the LR ensemble with LS SST and SSS biases of comparable magnitude to those of the HR ensemble (see Fig. A2 in the original manuscript). We would like to note, though, that results are significant if the whole LR ensemble size is considered, and that even in the case of a reduced LR sample size, the corresponding CI is clearly centered to the right of 0 (Table T1).

The reduction in the CNA SST and SSS biases observed in the HR ensemble mean is not significant, as the CI of the difference in means between the HR and LR ensembles does contain 0 (for all LR ensemble subsample sizes)(Table T1). We note though that CIs are notably centered to the right of 0. In the case of SSTs, the lack of significance is associated with the cold biases in MPI-ESM1-2-ER and EC-Eart3P-VHR still present in that area (Fig. A2). For SSS, the lack of significance is related to the fact that several LR models have a similar performance to the HR models in that region. Also, we note that the MPI-ESM1-2-ER model still presents a significant SSS bias in the CNA (Fig. A2).

As for the NCH region, the analysis shows that both SST and SSS biases are significantly reduced in the HR ensemble compared to the LR ensemble (Table T1). However, in the case of SSTs, samples of at least size 8 are required from the LR ensemble to achieve this significance, which is due to the warm bias that CESM1-CAM5-SE-HR presents in that area (Fig. A1).

Regarding max. AMOC at 26.5 ºN, although CIs are notably centered to the left of 0, the reduction in strength in the HR ensemble is not significant (Table T1), since several LR models show values within the range of the HR ensemble (Fig. 11a). Interestingly, the distance to the RAPID profile (as measured by the RMSE) is significantly reduced in the HR ensemble (for all LR ensemble subsample sizes; Table T1). The increase in correlation to the RAPID curve in the HR ensemble becomes significant when samples considered in the bootstrapping from the LR ensemble have a minimum size of 14, due to several LR models presenting correlation values in the same range of the HR ensemble (Fig. 11b).

Metric	**total LR ensemble size**	**reduced LR ensemble size**	**min. LR size for 95% sign.**
LS SST (ºC)	**[0.21 1.86]**	[-0.79 3.11]	19
LS SSS	**[0.04 1.00]**	[-0.39 1.78]	25
CNA SST (ºC)	[-1.05 3.07]	[-1.43 3.66]	-
CNA SSS	[-0.18 1.01]	[-0.63 1.70]	-
NCH SST (ºC)	**[-4.58 -0.16]**	[-5.19 0.23]	8
NCH SSS	**[-2.54 -1.02]**	**[-2.99 -0.38]**	4
max AMOC (Sv)	[-5.10 0.66]	[-7.75 2.61]	-
RMSE AMOC (Sv)	**[-1.59 -0.60]**	**[-2.69 -0.05]**	4
correl AMOC	**[0.01 0.11]**	[-0.02 0.20]	14
RMSE temp profile (ºC)	**[-1.31 -0.39]**	**[-1.57 -0.12]**	4
correl temp profile	[-0.15 0.16]	[-0.16 0.20]	-
RMSE salt profile	**[-0.41 -0.09]**	[-0.68 0.04]	8
correl salt profile	**[0.01 0.04]**	[-0.00 0.07]	6
RMSE density profile (kg m-3)	**[-0.26 -0.06]**	[-0.48 0.03]	8
correl density profile	**[0.01 0.04]**	[-0.00 0.06]	7
max MLD (m)	[-310.52 1530.64]	[-703.04 1858.65]	-
max SPG (Sv)	[-11.17 9.94]	[-17.10 14.09]	-

Table T1: The first column indicates the single numeric metrics analyzed (units in parenthesis). The second column shows the 95% CI of the differences in means between the HR and LR ensembles, calculated from a distribution of bootstrapping samples with repetition. The size of the samples coincides with the total size of their respective ensembles. The third column is analogous to the second one but in this case the size of the LR samples coincides with the total size of the HR ensemble. The fourth column indicates the minimum size of the LR samples in the bootstrapping required to obtain a CI not containing the value 0. Text in bold indicates when this is the case.

In terms of temperature profiles in the LIS box, RMSEs relative to EN4 are significantly reduced in the HR ensemble compared to LR, even when considering small subsamples of size four in the bootstrapping analysis. The reduction in RMSEs is particularly pronounced for the EC-Earth3P-VHR and MPI-ESM1-2-ER models (Fig. 6a). The increase in correlation with respect to the EN4 temperature profile in the HR ensemble is not significant, which is due to the low correlation exhibited by HadGEM3-GC31-HH (Fig. 6a). By removing this model from the bootstrapping calculations, correlation becomes significant even with a reduced LR subsample size (not shown). Regarding the salinity and density profiles, improvements in the HR ensemble related to both RMSE and correlation to EN4 become significant already with relatively small LR sample sizes (Table T1).

The increase in max. MLD observed in the HR ensemble compared to LR is not significant (Table T1), since several LR models present max. MLD within the same range displayed in the HR ensemble (Fig. 7). We note though that CIs are again centered well to the right of 0.

The increase in the SPG strength in the HR ensemble is also not significant (Table T1), again because several LR models present values within the same range as the HR ensemble (Fig. 12).

The manuscript will be edited to reflect all the findings described in this point.

My other concern is that the analysis here is quite basic and doesn’t really show much that is new. The authors could include more analysis of scatterplots of the metrics they analyse against each other and discuss the implications for how biases affect each other. There have been a number of studies looking at how resolution affects the North Atlantic. The novelty of this paper seems to be having multiple models at eddy-resolving rather than permitting resolution. What are the implications of going to eddy-resolving resolution?

We have added scatterplots relating some of the metrics in Table T1, as well as a discussion of the associated findings in the updated version of this manuscript. More specifically, the new analyses include scatterplots of modelled Labrador Sea SSS biases versus AMOC strength, mixed layer depth, and SPG strength, as well as scatterplots of mixed layer depth vs AMOC strength, and vs SPG strength (please, see Figure F1 below as an example).
Some of the implications of going to eddy-resolving resolution have already been discussed in the previous point about bootstrapping.

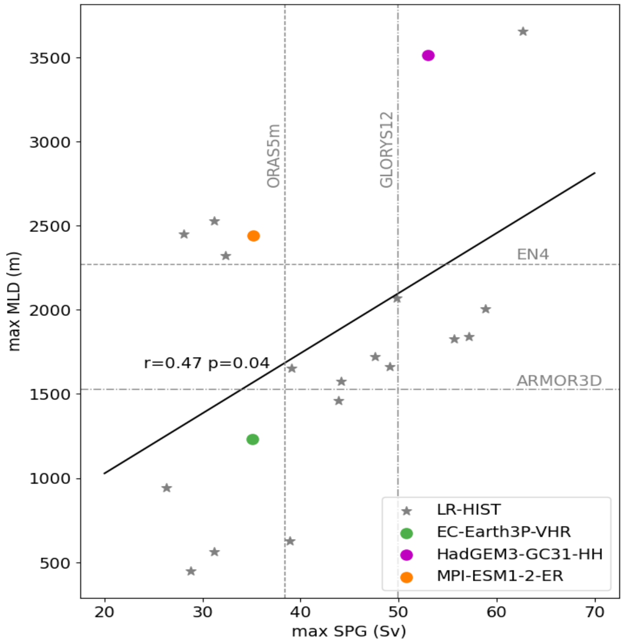

Figure F1. Scatterplot of max. SPG strength (in Sv) vs max. MLD (in m), both referred to the LIS box (shown in Fig. 8). Pearson correlation coefficient and p-value are shown next to the fit line. Horizontal dashed and dot-dashed lines show EN4 and ARMOR3D observation-based values, respectively. Vertical dashed and dot-dashed lines show ORAS5m and GLORYS12 reanalysis values, respectively.

Minor

-L11 ‘important role in featuring global ocean dynamics’ – what does this mean?

This sentence has been rephrased for clarification and the impact on modelled climate of resolving ocean mesoscale structures is discussed in the introduction (old lines 56-70).

New text: “Ocean mesoscale processes, which are parameterized in models with standard resolutions on the order of 1º or coarser, have an impact at larger scales, affecting the ocean mean state and circulation.”

Old text: “Ocean mesoscale structures, which are parameterized in standard resolution models, play an important role in featuring global ocean dynamics.”

-L19 ‘weaker than for lower resolution models’ This is rather misleading – the way this is reported in the abstract suggests that it is a result of the resolution change. As the authors discuss, studies have shown that the impact of increased resolution varies from model to model. The result here that the AMOC is weaker in the LR ensemble is likely because of some very strong models in the LR ensemble.

We have removed this statement.

New text: “the Atlantic Meridional Overturning Circulation (AMOC) is closer to RAPID observations.”

Old text: “the Atlantic Meridional Overturning Circulation (AMOC) is weaker than for lower resolution models and closer to RAPID observations”

-L63 ‘to a’-> ‘with a’

Changed, also at L64.

-L105 Include resolution in km to compare with other values.

Done.

New text: “of at least 1/10º (10 km).”

-L173 I can’t see the paper ‘in review’ though other studies (e.g. Marzocchi et al 2015) have suggested that this bias is because of the location of the NAC, rather than the strength.

We agree with the reviewer that also the location of the NAC plays a role in the magnitude of the CNA bias. We have edited the text accordingly. We would like to point out that Fig. 8 in Marzocchi et al., 2015 shows larger velocities in the North Atlantic at increased ocean resolution, including in the CNA region, and states that "increasing the model's resolution leads to an overall increase of absolute velocities, not only for western boundary currents but also for the entire domain…". Lin et al. is still under review, and would be removed from the text if Frigola et al. gets accepted and Lin at al. has not been accepted yet.

New text: "The other is a cold bias in the CNA (2–5º C), which earlier studies have linked to an unrealistic position of an overly weak NAC (Marzocchi et al., 2015) and an underestimation of the horizontal heat transport into the CNA domain (Lin et al., in review)."

Old text: "The other is a cold bias in the CNA (2–5º C), which earlier studies have linked to an overly weak NAC and an underestimation of the horizontal heat transport into the CNA domain (Lin et al., in review)."

-L237 and Fig 6 Using correlation of profiles to assess the shape seems rather flawed – if one profile has twice the slope of a second profile then the correlation would be perfect. It would be better to assess the stratification itself.

We agree that the use of correlation coefficients as a stand-alone metric to assess the resemblance between two different vertical profiles should be avoided, for the reason stated by the reviewer, i.e., one curve could have point-to-point slopes that are a (large) multiple of the other curve's slopes. Nevertheless, when a high correlation between two curves is observed in conjunction with a small RMSE, this is an indication that we are not in that case, i.e., that the slopes of the two curves are similar and that the distance between curves is small. That's the reason we considered both metrics together.

We have edited our text to ensure that correlation coefficients are not used as a stand-alone metric in our statements (see below). Additionally, we have added this text at the end of line 240:

"We note that the use of vertical correlation coefficients to assess resemblance between two vertical profiles should come in conjunction with other metrics, such as RMSEs, or direct visual inspection of profiles, as a high correlation coefficient alone does not ensure a small distance between curves."

Lines 236-238 have been rephrased as:

New text: "Indeed, the density profile for HR-HIST is closer in shape to the EN4-derived one, as supported by the Pearson correlation coefficients in Fig. 6c, which are very close to one in all HR-HIST models, and by the relatively small RMSEs."

Old text: "Indeed, the density profile for HR-HIST is closer in shape to the EN4-derived one, as supported by Pearson correlation coefficients in Fig. 6c, which are very close to one in all HR-HIST models.

Additionally, we have slightly edited Lines 227-229:

New text: “Overall, the vertical salinity profile exhibits a more realistic shape in HR-HIST, a higher correlation coefficient, and a slightly smaller RMSE against EN4 (Figs. 5b, 6b).”

Old text: “Overall, the vertical salinity profile is more realistic in HR-HIST, as supported by the higher correlation coefficient and smaller RMSE against EN4 indicated in Fig. 6b.”

-L243 ‘expected to impact on’

Changed.

-Fig 6 What region are these for?

The Pearson correlation coefficients and RMSEs are associated to the vertical profiles in Fig. 5, which are averaged over the region (35º–60º W, 50º–65º N), shown in Figs. 7 and 8 (blue box). The captions in Fig. 6 have been edited for clarification.

New text:

Figure 6. Pearson correlation coefficient (horizontal axis) and Root Mean Square Error (RMSE) (vertical axis; units as in Fig. 5) for the vertical (a) temperature, (b) salinity and (c) density profiles in Fig. 5 against EN4. Profiles are averaged over the region (35º–60º W, 50º–65º N), the LIS box shown in Figs. 7 and 8.

Old text:

Figure 6. Pearson correlation coefficient (horizontal axis; units as in Fig. 5) and Root Mean Square Error (RMSE) (vertical axis) of the (a) temperature, (b) salinity and (c) density profiles in Fig. 5 against EN4, in the vertical dimension.

-L256 ‘Its’ -> ‘The’

Changed.

-Fig 8 Might be clearer if you adjusted the scale

Done.

-L287 I don’t really understand this sentence.

The better agreement of the AMOC between the HR-HIST ensemble mean and ORAS5m (compared to the AMOC between the LR-HIST ensemble mean and ORAS5m) might not only be due to a more realistic representation of the AMOC in HR-HIST. The fact that ORAS5m AMOC data were produced with an eddy-permitting version of the ocean model NEMO, which in this sense is closer to the models in the HR-HIST ensemble (eddy-resolving), than to the ones in the LR-HIST ensemble (eddy-parameterized), might also contribute to the resemblance of the HR-HIST ensemble mean and ORAS5m AMOC representations, as part of the AMOC variability will by driven by the model physics.

The old text has been rephrased for clarification.

New text:

“The larger resemblance of the AMOC streamfunction in ORAS5m and the HR-HIST ensemble mean (compared to the LR-HIST ensemble mean) might be to some extent related to the fact that ORAS5m is run with an eddy-permitting model, and is thus potentially more similar to the models in the HR-HIST ensemble (eddy-resolving), than to the models in the LR-HIST ensemble (eddy-parameterized).”

Old text:

“This improved agreement with ORAS5m can partly be attributed to the use of an eddy-permitting ocean model in ORAS5m, allowing for the representation of some eddies.”

-L305 It’s not entirely clear what you mean by methodological approach and what it affects – expand a little on this.

Details on the methodological approaches have been added to the text. Please, see below.

New text:

” The HR-HIST mean profile shows a particularly good fit with the RAPID array one above ~1000 m, although, in general, the AMOC streamfunction is too shallow both for LR-HIST and HR-HIST. Some differences between models and observations might stem from the methodologies used to derive the AMOC profiles. While in models the AMOC is obtained by integrating model velocities, which are simulated in every grid point, this approach is not possible with observations, since direct velocity measurements are scarce. RAPID data combines measurements of four separate AMOC components. The first is the Florida Current transport, which has been inferred from cable voltage measurements west of the Bahamas since 1982 (Larsen and Sandford, 1985). The second one is the western boundary wedge transport, which measures elements of the Antilles and the deep western boundary currents using current meters west of 76.75ºW to Abaco Island. The third term is the near-surface AMOC Ekman transport, calculated itself from wind stress reanalysis data. The fourth term is the upper mid-ocean return transport, derived for the region east of 76.75º W from density profiles using zonal gradient of dynamic heights (Roberts et al., 2013; Danabasoglu et al., 2021; McCarthy et al., 2015). The calculation of this gradient in RAPID makes use of a reference depth (4820 m), which represents a level-of-no-motion. Some studies report sensitivity of the estimated RAPID profile, particularly in the deep ocean, to the choice of this reference depth (Fig 3.2 in McCarthy et al., 2015; Fig. S3 in Roberts et al, 2013), which might explain some of the differences between the RAPID and model profiles in the deep ocean (Fig. 11). However, uncertainties related to the choice of a reference depth are within the range of the accuracy of the RAPID method, and uncertainties in deep transport are a current topic in the literature (McCarthy et al., 2015). A model-based study also suggests that estimating the AMOC via RAPID's physical assumptions could lead to an underestimation of up to 1.5 Sv in its mean value at ~900 m depth (Sinha et al., 2018) compared to its real strength, a result that is, however, not supported in a more recent study based on a different ocean model (Danabasoglu et al. 2021)."

Old text:
"The HR-HIST mean profile shows a particularly good fit with the RAPID array one above ~1000 m, although, in general, the AMOC is too shallow both for LR-HIST and HR-HIST. This is in part due to differences in the methodological approach (Danabasoglu et al., 2021)."

-L331 'biases in NCH'

Added.

-L347 'reduced cold and fresh biases'

Changed.

-L380-383 and Fig A1 Needs more discussion. The correlations in Fig 1a are certainly not significant so I'm not sure what it means that the correlation increases. Maybe instead discuss why there might be correlations and why the HR models might be outliers.

The old text has been edited and expanded as follows.

New text:

"Since the NCH region falls within the GS domain in LR-HIST models, we might expect the NCH and CNA biases to be correlated for those models, due to their ultimate link with the GS/NAC dynamics. Indeed, SSSs between the two regions show a significant correlation if we restrict our analysis to the LR-HIST ensemble (Fig. A1b). Interestingly, when HR-HIST models are included in the calculations, the SSS correlation between the NCH and CNA decreases/loses its significance (p=0.06), which we argue is due to the fact that NCH is outside the GS domain in HR-HIST models. The correlation in SST temperature between the NCH and CNA regions is not significant for LR-HIST models though, which could be related to some damping of the SST signal through interactions with the atmosphere (Fig. A1a). We note that HR-HIST models tend to appear on the lower side of the point cloud in Fig. A1, which reflects the reduced warm and salty biases in the NCH region in HR-HIST models".

Old text:

"Surface biases in the NCH and CNA regions might be connected with each other due to their ultimate link with the GS/NAC dynamics. As expected, the correlation between the surface biases of these two regions increases when the HR-HIST models are excluded from the computation (Fig. A1), which reflects that the NCH region is outside the GS domain in HR-HIST models."

-L384-386 Doesn't really make sense – what is 'it'?

We have rephrased the text. 'It" referred to "the bias".

New text:" In the HR-HIST multi-model mean, the LS region stands out for a warm and salty bias. The LR-HIST ensemble mean shows also a warm bias in the LS, although this is much weaker in magnitude compared to the one in the HR-HIST ensemble mean. No salty bias is present in the LS in the LR-HIST ensemble mean, although some individual LR-HIST models do show a salty bias in that region, comparable in magnitude to that of the HR-HIST ensemble (Fig. A2)".

Old text: "The LS region stands out for a warm and salty bias in the HR-HIST multi-model mean. For temperature, it is also present in the LR-HIST ensemble mean, although much weaker, whereas for salinity, it does not show in the LR-HIST ensemble mean, even if it is present in some of the individual models."

-L397 temperature signals are often damped by interactions with the atmosphere reducing the amplitude of the signal. This would explain why there are stronger correlations for salinity.

The sentence has been rephrased to include the reviewer's suggestion.

New text:

"The correlation between the SST biases of the LS and CNA regions is also significant, although weaker compared to the SSS biases (r = 0.54, p < 0.001; Fig. A2), which could be related to a damping of the SST signal through interactions with the atmosphere (which are usually more important than for salinity), or to mixing with Arctic waters."

Old text:

"The correlation between the SST biases of the LS and CNA regions is also significant, although weaker compared to the SSS biases (r = 0.54, p < 0.001; Fig. A2), which might indicate additional differences between the mechanisms exerting control over the SSTs in both regions, like the local atmospheric forcing."

-L401-404 I don't understand what the authors are getting at here – please explain more.

The text has been rephrased and extended.

New text:

"We find improved temperature and salinity stratification in the LIS box for HR-HIST compared to LR-HIST, with the HR-HIST ensemble mean curve closer in distance and shape to EN4. The warm and salty biases present at the subsurface in both ensembles over most of the column (below ~150m) are reduced in the HR-HIST ensemble mean. On the other hand, surface biases (above ~150 m) are more pronounced in the HR-HIST ensemble mean compared to the LR-HIST ensemble mean. The increased (reduced) biases at the surface (subsurface) for HR-HIST might be related to the fact that ocean mesoscale eddies increase vertical (upwards) heat and salt transports in the ocean (Hewitt et al., 2017). Although vertical transport by eddies provides a plausible explanation for the warm and salty biases in the LS, additional factors might be needed to correct for those biases."

Old text:

"Interestingly, we find improved vertical profiles of temperature and salinity in the LIS box for HR-HIST: despite the larger biases found at the surface with respect to LR-HIST, the subsurface is colder and fresher in HR-HIST. This might be related to increased vertical (upwards) heat and salt transports by ocean mesoscale eddies (Hewitt et al., 2017), providing a potential explanation for the warm and salty surface biases in the LS."

-L410-418 Different definitions of MLD can give very different results. Do all the studies mentioned use the same density criteria? Also you don't make it very clear that using monthly mean densities to calculate a MLD can give very different (and likely shallower) estimates than means of instantaneous profiles used by Yashayaev etc.

Yashayaev and Loder (2016; 2017), Holte et al. (2017), and our study, all use a different methodology (a description of the different methods is provided below in the new text). We believe that the exact effect of combined temporal and spatial data smoothing in the different studies might be difficult to assess. Nevertheless, we have highlighted the different temporal and spatial characteristics of the data in the studies as one of the causes for the differences in the range of MLD estimates.

New text:

"A range of methods is used in the different studies. The yearly estimates by Yashayaev and Loder (2016; 2017) are winter maximum values of "aggregate" maximum convection depths, defined as the 75th percentile of the depth of the base of the pycnostad in the set of available individual LS profiles at each time. To estimate the depth of the pycnostad for each profile, first, layer thicknesses of σ_1 potential density classes (binned in 0.005 kg m^{-3} intervals) are calculated (Fig. 3b in Yashayaev and Loder, 2016); subsequently, the lower boundary of each pycnostad is defined as the depth corresponding to the σ_1 value with the largest layer thickness (plus a constant). Holte et al. (2017)'s MLD estimates (shown in their Fig. 3a) correspond to individual Argo profiles and are obtained with a density algorithm (Holte and Talley, 2009) that uses temperature, salinity, and density data from individual profiles to calculate MLD through a combination of methods and elements, including temperature and density threshold methods (with threshold values of 0.2ºC and 0.03 kg m^{-3}, respectively), temperature and density gradient methods, maximum/minimum values of temperature, salinity and density over the profiles, estimates of thermocline linear fits, etc. Meanwhile, in our study, MLD values are a climatology of March MLD monthly means obtained from gridded temperature and salinity data through a density threshold method based on monthly data (threshold value = 0.03 kg m^{-3}). Overall, the maximum MLD values in our HR-HIST ensemble mean for the LS (1800–2000 m) are larger than the observational estimates, excluding the record values in Yashayaev and Loder (2017). These differences might arise from the different methodological approaches, from the different temporal and spatial characteristics of the profile data, as well as from differences in the time intervals analyzed."

Old text:

"Overall, the maximum values in the HR-HIST ensemble mean for the LS (1800–2000 m) are slightly larger than the direct observational estimates, excluding the record values in Yashayaev and Loder (2017). However, it is important to note that some differences are expected as our HR-HIST values are computed from 1980 to 2014, and use generally smoother profiles associated with the coarser temporal resolution of the model data compared to the individual profiles from observational studies. "

-L463 'southward propagation of the MLD signal into the AMOC' – what does this mean?

We have expanded the original text. The results described are visible in Fig. 7 in Martin-Martinez et al. (2024).

New text:

"Recent work by Martin-Martinez et al. (2024) shows that ocean resolution also affects the timescales governing large-scale dynamical processes in the North Atlantic. In that study, Labrador Sea mixed layer depth is found to be positively correlated with overturning streamfunction strength at high latitudes at 0 time lags in all resolutions of the HighResMIP EC-Earth3P-VHR control simulations. This imprinted signal of the mixed layer in the overturning streamfunction at high latitudes propagates to lower latitudes at subsequent lags. Interestingly, the propagation speed is significantly larger in eddy-resolving models, compared to coarser resolution models (Fig. 7 in Martin-Martinez et al., 2024)"

Old text:

"Results by Martin-Martinez et al. show a faster southward propagation of the MLD signal into the AMOC in eddy-resolving models, indicating that ocean resolution affects also the timescales of the dynamics of the North Atlantic."

References:

Danabasoglu, G., Castruccio, F. S., Small, R. J., Tomas, R., Frajka-Williams, E., and Lankhorst, M.: Revisiting AMOC Transport Estimates From Observations and Models, Geophysical Research Letters, 48, e2021GL093045, https://doi.org/10.1029/2021GL093045, 2021.

Hewitt, H. T., Bell, M. J., Chassignet, E. P., Czaja, A., Ferreira, D., Griffies, S. M., Hyder, P., McClean, J. L., New, A. L., and Roberts, M. J.: Will high-resolution global ocean models benefit coupled predictions on short-range to climate timescales?, Ocean Modelling, 120, 120–136, https://doi.org/10.1016/j.ocemod.2017.11.002, 2017.

Holte, J. and Talley, L.: A New Algorithm for Finding Mixed Layer Depths with Applications to Argo Data and Subantarctic Mode Water Formation, https://doi.org/10.1175/2009JTECHO543.1, 2009.

Holte, J., Talley, L. D., Gilson, J., and Roemmich, D.: An Argo mixed layer climatology and database, Geophysical Research Letters, 44, 5618–5626, https://doi.org/10.1002/2017GL073426, 2017.

Larsen, J. C. and Sanford, T. B.: Florida Current Volume Transports from Voltage Measurements, Science, 227, 302–304, https://doi.org/10.1126/science.227.4684.302, 1985.

Martin-Martinez, E., Frigola, A., Moreno-Chamarro, E., Kuznetsova, D., Loosveldt-Tomas, S., Samsó Cabré, M., Bretonnière, P.-A., and Ortega, P.: Effect of horizontal resolution in North Atlantic mixing and ocean circulation in the EC-Earth3P HighResMIP simulations, EGUsphere, 1–33, https://doi.org/10.5194/egusphere-2024-3625, 2024.

Marzocchi, A., Hirschi, J. J.-M., Holliday, N. P., Cunningham, S. A., Blaker, A. T., and Coward, A. C.: The North Atlantic subpolar circulation in an eddy-resolving global ocean model, Journal of Marine Systems, 142, 126–143, https://doi.org/10.1016/j.jmarsys.2014.10.007, 2015.

McCarthy, G. D., Smeed, D. A., Johns, W. E., Frajka-Williams, E., Moat, B. I., Rayner, D., Baringer, M. O., Meinen, C. S., Collins, J., and Bryden, H. L.: Measuring the Atlantic Meridional Overturning Circulation at 26°N, Progress in Oceanography, 130, 91–111, https://doi.org/10.1016/j.pocean.2014.10.006, 2015.

Roberts, C. D., Waters, J., Peterson, K. A., Palmer, M. D., McCarthy, G. D., Frajka-Williams, E., Haines, K., Lea, D. J., Martin, M. J., Storkey, D., Blockley, E. W., and Zuo, H.: Atmosphere drives recent interannual variability of the Atlantic meridional overturning circulation at 26.5°N, Geophysical Research Letters, 40, 5164–5170, https://doi.org/10.1002/grl.50930, 2013.

Sinha, B., Smeed, D. A., McCarthy, G., Moat, B. I., Josey, S. A., Hirschi, J. J.-M., Frajka-Williams, E., Blaker, A. T., Rayner, D., and Madec, G.: The accuracy of estimates of the overturning circulation from basin-wide mooring arrays, Progress in Oceanography, 160, 101–123, https://doi.org/10.1016/j.pocean.2017.12.001, 2018.

Yashayaev, I. and Loder, J. W.: Recurrent replenishment of Labrador Sea Water and associated decadal-scale variability, Journal of Geophysical Research: Oceans, 121, 8095–8114, https://doi.org/10.1002/2016JC012046, 2016.

Yashayaev, I. and Loder, J. W.: Further intensification of deep convection in the Labrador Sea in 2016, Geophysical Research Letters, 44, 1429–1438, https://doi.org/10.1002/2016GL071668, 2017.