Summary

Frigola et al. investigate the role of resolving mesoscale ocean eddies on the representation of the mean state & circulation of the North Atlantic Ocean in coupled climate models. The authors present a largely qualitative comparison of four coupled historical simulations with nominal ocean resolutions of at least 1/10° to an ensemble of 39 coupled simulations configured at coarser horizontal resolution. The study concludes that the vertical stratification & deep convection in the subpolar North Atlantic, and both the meridional overturning and barotropic circulations agree more closely with ocean observations at mesoscale eddy-resolving resolution. The manuscript is generally well written and includes a valuable final discussion; however, I have significant concerns regarding:

(1) the use of the phrase 'eddy-resolving',

(2) the ocean observations and reanalysis products used to evaluate model performance, and

(3) the originality of the study & its wider implications.

I would not recommend the manuscript for publication until major revisions have been made to address each of the comments below.

We would like to sincerely acknowledge the reviewer for their constructive comments that have greatly contributed to improving this manuscript. The above-mentioned points have been addressed. The use of the phrase "eddy-resolving" has been refined to "mesoscale eddy-resolving" throughout the manuscript, new observations and reanalysis have been included as suggested, as well as new diagnostics exploring inter-model differences and relationships between the different variables analyzed (please, see below for details).

General Comments

- Use of 'eddy-resolving':

In both the title and throughout the manuscript, the authors use 'eddy-resolving' to refer to coupled model simulations with sufficiently fine horizontal ocean resolution to resolve mesoscale eddies. Given the recent emergence of both submesoscale permitting / resolving ocean model configurations (e.g., Chassignet & Xu, 2017; Lévy et al., 2010; Pennelly & Myers, 2020; Pennelly & Myers, 2022; Li et al., 2023) and a growing awareness that representing or resolving submesoscale processes is integral for the accurate simulation of North Atlantic mean state (see Jackson et al, 2023 for a review), I would strongly recommend that the authors refine their use of 'eddy-resolving' to 'mesoscale resolving' throughout.

Done. We have rephrased "eddy-resolving" as "mesoscale eddy-resolving" throughout the manuscript.

Similarly, I would suggest revising the manuscript title to: 'The North Atlantic Ocean mean state in mesoscale eddy-resolving coupled models: a multi-model study' or equivalent.

Done.

New title: 'The North Atlantic Ocean mean state in mesoscale eddy-resolving coupled models: a multi-model study'.

Furthermore, the authors should explicitly address the role of submesoscale features and the implications of their (justifiable) absence in current generation coupled climate models for the North Atlantic mean state; for example, their important role in restratifying the Labrador Sea (Clément et al., 2023; Frajka-Williams et al., 2014) and reducing deep convection (e.g., Tagklis et al. (2020)) and Gulf Stream penetration (e.g., Chassignet and Xu (2017) and Chassignet et al. (2020)).

Thanks for this remark. We have added a paragraph in the introduction (right before old line 71) discussing the contribution of submesoscale eddies to the North Atlantic mean state and the implications of their absence.

New text:

"Also submesoscale processes have an impact on the North Atlantic mean state. Tagklis et al. (2020) show a significant reduction in deep water convection in the Labrador Sea (and an increase in vorticity) when increasing the grid resolution in a regional model from 15 km (mesoscale permitting in the Labrador Sea) to 1 km (submesoscale resolving in the Labrador Sea). That study finds that the simulated reduction in convection is caused by eddy heat advection from the Irminger Current and by local submesoscale eddy buoyancy fluxes from the Labrador Sea basin itself. Similarly, restratification of the Labrador Sea convective areas at the end of winter has been associated with both mesoscale and submesoscale eddies (Clément et al., 2023). Another example of the importance of the submesoscale in the representation of the North Atlantic dynamics is a further eastward penetration of the NAC and its eddy variability at 1/50º resolution, in closer agreement with observations, compared to mesoscale resolving scales (Chassignet and Xu, 2017). Omitting submesoscale eddies contributions might thus imply biases in the representation of the NAC and deep water convection. Current computational resources allow for multidecadal global coupled runs at mesoscale resolving resolutions, so that future research efforts should be aimed at parameterizing submesoscale-related processes to the extent possible."

The manuscript would also benefit from greater effort to contextualise inter-model differences within the LR & HR ensembles; for example, do all HR simulations use the same z-level vertical coordinate system & what impact would any difference have on the representation of the overflows (e.g., Colombo et al., 2020, Bruciaferri et al., 2024) and Labrador Sea stratification downstream (MacGilchrist et al. 2020).

In order to help identify and explain inter-model differences, we have added and discussed new figures (which can be found in the updated manuscript) relating the different variables analyzed in our study. 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 F6 below as an example).

Regarding the vertical coordinate system, EC-Earth3P-VHR and HadGEM3-GC31-HH share the same z-level vertical coordinates, with a total of 75 depth levels. CESM1-CAM5-SE-HR uses a different z-level vertical grid and it has 62 depth levels. MPI-ESM1-2-ER has the coarsest vertical grid of all, with 40 levels. As for the horizontal resolutions, the finest are again those of HadGEM3-

GC31-HH and EC-Earth3P-VHR (1/12º), followed by CESM1-CAM5-SE-HR and MPI-ESM1-2-ER (1/10º).
Ocean models using fixed vertical levels (z-models) present difficulties in correctly representing the densities of the Arctic overflows downslope the Greenland-Scotland Ridge (Bruciaferri et al., 2024; Colombo et al., 2020; Jackson et al., 2023), which in turn affects stratification in the subpolar North Atlantic and transport in the AMOC lower limb (Bruciaferri et al., 2024).
Results by Colombo et al. (2020) suggest that increasing horizontal resolution to submesoscale scales in combination with an increase in vertical resolution improves the representation of the Denmark Strait overflow, although at a fixed horizontal resolution of 1/12º, increasing the vertical resolution alone does not lead to any improvement (Colombo et al. 2020). If the properties of the overflows were exclusively based on the model resolution and not dependent on the model physics, we could hypothesize that the EC-Earth3P-VHR and HadGEM3-GC31-HH models might have a better representation of the overflows compared to MPI-ESM1-2-ER and CESM1-CAM5-SE-HR, due to their slightly finer combined horizontal and vertical resolutions. However, a detailed analysis of the differences in the representation of the overflows in the HR-HIST models was beyond the scope of this study.

Nonetheless, we have added a paragraph in the discussion section of the manuscript explaining the challenges presented in z-models in representing the overflows (as described above) and how the introduction of local terrain-following coordinates near the Greenland-Scotland Ridge in models could represent a source of improvement (Bruciaferri et al., 2024).

- Datasets used for model validation:

My second major concern is both the choice of ocean observations and reanalysis datasets used to validate the ocean model components & the details absent from their methodology descriptions.

We have addressed each of these points right below.

More specifically, I could not find justification for why a coarse resolution ocean analysis product (EN4.2.2) is used to validate mesoscale-resolving ocean models, when at least eddy-permitting resolution products are available (e.g., ARMOR3D [https://doi.org/10.48670/moi-00052] or ASTE [Nguyen et al., 2021]). While no observational product is a 'true' representation of reality, I would argue it is more appropriate to compare the property fields simulated in mesoscale resolving models with observational products that can, at least partially, represent them.

Thanks for this remark. The ARMOR3D dataset at 0.25º resolution (Guinehut et al. 2012), which we were not aware of, has now been added to our analysis of the MLD to complement the previous EN4 estimates and to provide further robustness to our results (please, see Fig. F1). We would like to note, though, that the EN4 1º data have been previously employed to validate both model MLD fields at eddy-permitting and eddy-resolving resolutions (Koenigk et al., 2021; Martin-Martinez et al., 2024), and model temperature and salinity fields in mesoscale eddy-resolving models (Jackson et al. 2023; Chassignet et al. 2020; Roberts et al. 2020; Gutjahr et al. 2019; Roberts et al. 2019; Moreno-Chamarro et al. 2025; and Martin-Martinez et al. 2024 use EN4; Marzocchi et al. 2015 use EN3). For that reason, we have kept it for better comparability with those previous studies. Besides, EN4 data span the entire historical period covered in our analysis (i.e. 1980-2014), making it especially suitable for bias assessment in the historical period, meanwhile ARMOR3D data only covers from 1993 onwards.

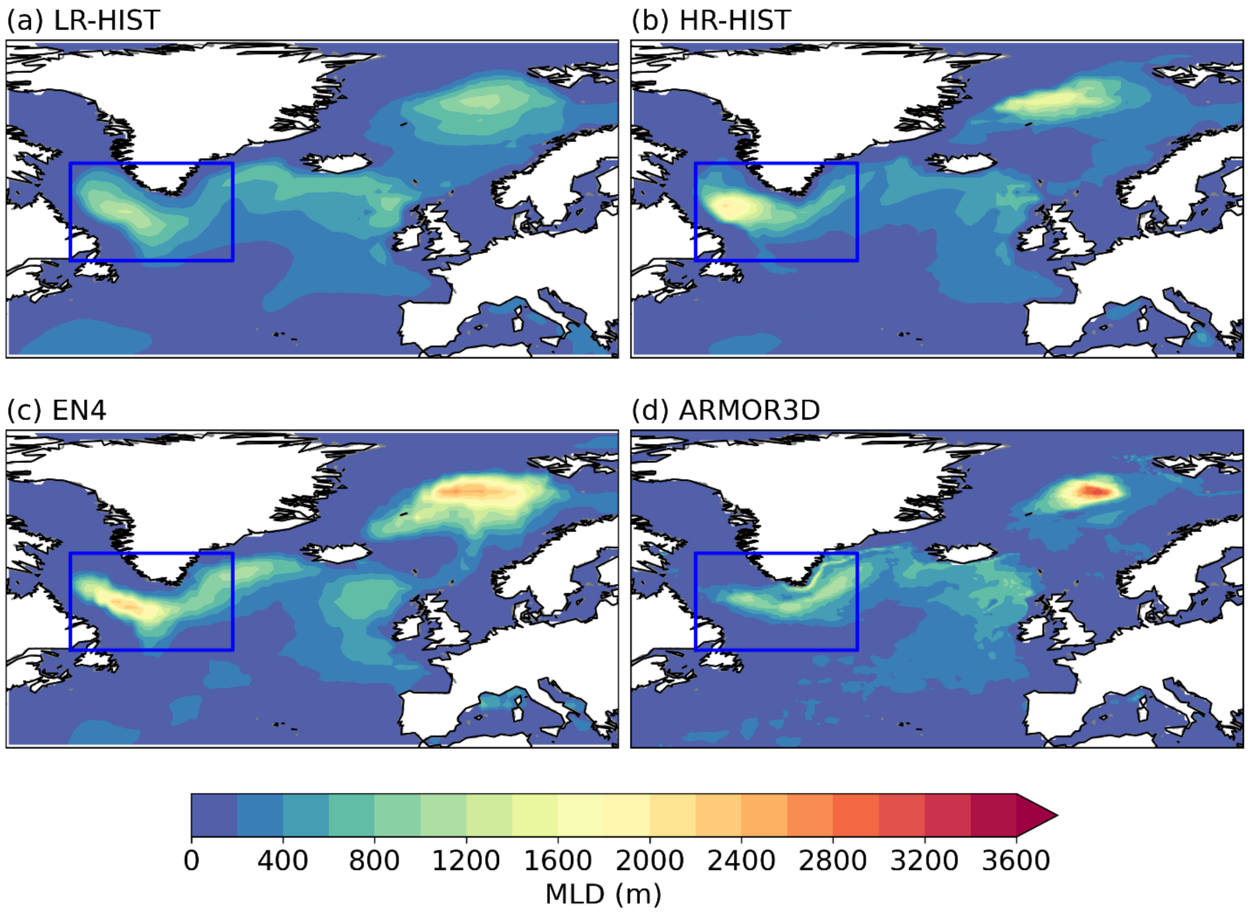

Figure F1. March MLD (in m) by groups, for (a) LR-HIST, (b) HR-HIST, (c) EN4, and (d) ARMOR3D. In all cases, including ARMOR3D, MLD has been calculated from temperature and salinity fields using the density threshold method of 0.03 kg m^{-3} described in the manuscript. The time interval covered is 1980-2014 in (a), (b), (c), and 1993-2014 in (d).

We would like to note a characteristic feature of the ARMOR3D dataset, namely a distinct stripe of deep mixing attached to the shelf along the East Greenland Current (Fig. F1).

Similarly, the authors could have used the mesoscale-resolving GLORYS12 ocean reanalysis product (https://doi.org/10.48670/moi-00021 - on its original NEMO grid) rather than the eddy-permitting ORAS5m reanalysis product to validate the mean meridional overturning stream function in depth space.

The GLORYS12 overturning and barotropic streamfunctions have been added to our analysis for further robustness of the results, to complement the ORAS5m dataset (0.25º resolution)(please, see Figs. F2 and F3). We have calculated both streamfunctions using GLORYS12 velocity fields (https://doi.org/10.48670/moi-00021), which are provided on a regular grid at 1/12° (0.083°).

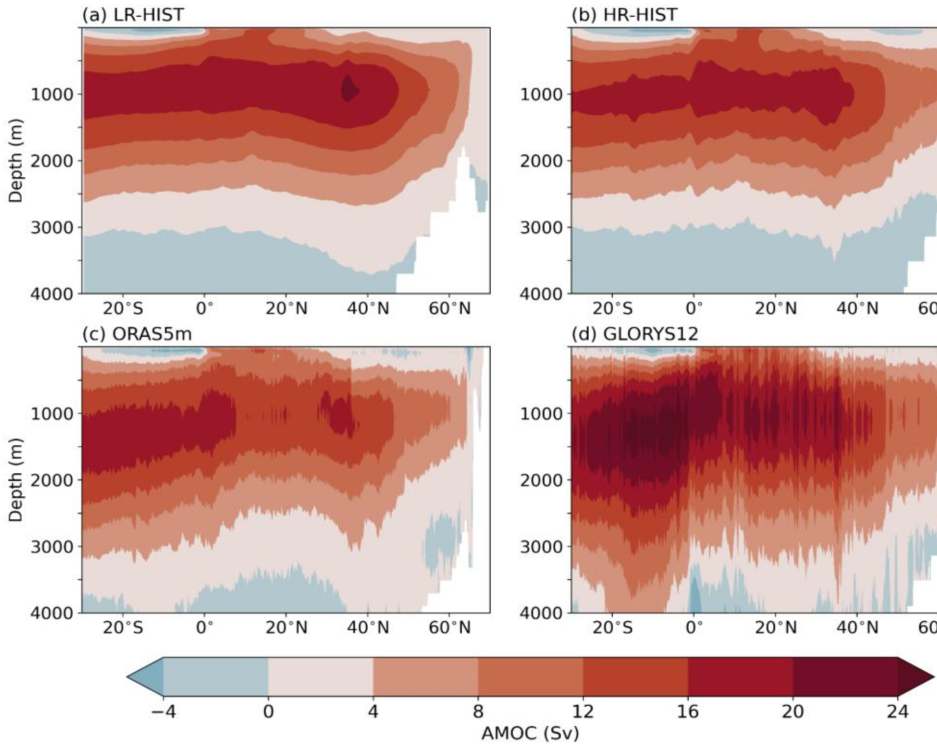

Figure F2. AMOC streamfunction (in Sv) by groups, for (a) LR-HIST, (b) HR-HIST, (c) ORAS5m, and (d) GLORYS12. The time interval covered is 1980-2014 in (a), (b), (c), and 1993-2014 in (d).

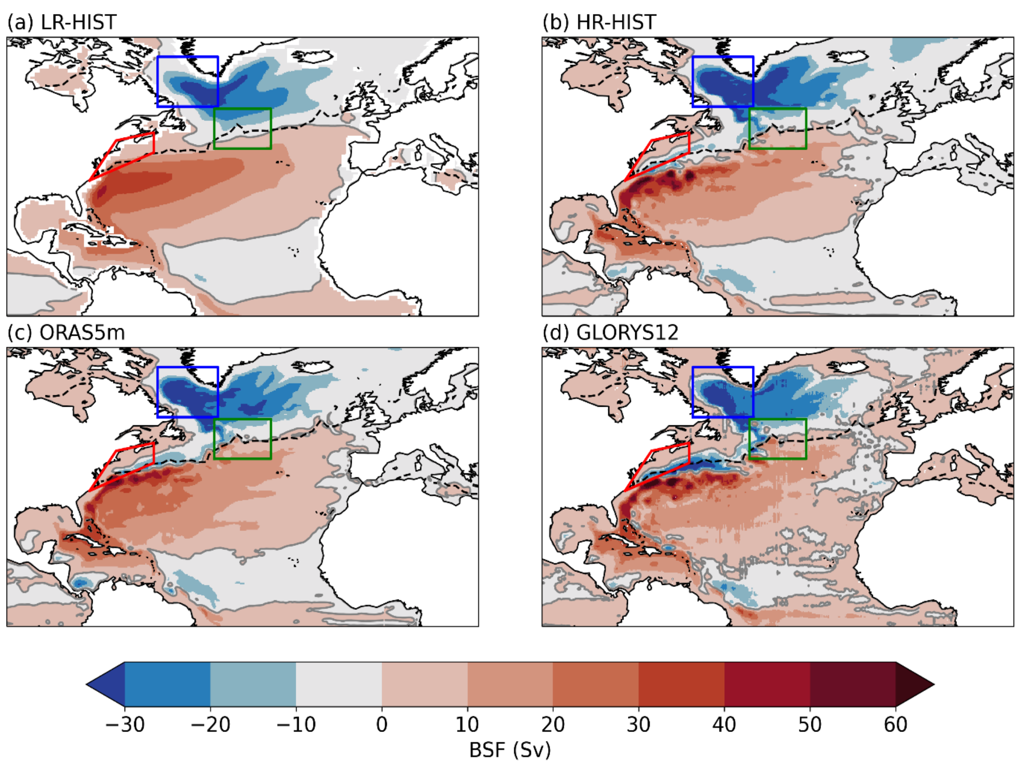

Figure F3. BSF (in Sv) by groups for (a) LR-HIST, (b) HR-HIST, (c) ORAS5m and (d) GLORYS12. The time interval covered is 1980-2014 in (a), (b), (c), and 1993-2014 in (d).

There are also some important details missing from Section 2.3 of the methodology; for example, are model property fields regridded onto the observations or vice versa for validation? And what type of interpolation is used: bilinear, conservative etc? The current use of interpolating colour contour plots does not make this obvious to readers.

For direct visual comparison of mean-state properties across models (e.g. in Figs. 7, 9, 12) data are plotted on the original grid. When some computation between models or against observations is required (e.g., multi-model means, model biases), model data are mostly regridded onto the observation's grid or, in some cases, onto a common regular grid. Methods employed are linear/bilinear (according to the dimension of the data), and nearest neighbour interpolation (closest source point), depending on the case. In all cases, a comparison between the regridded and the original data was performed, to ensure suitability of the interpolation scheme.
More specifically, in Figs. 1-4 model data are regridded to the EN4 regular grid using nearest neighbour interpolation (closest source point). In Fig. 5 individual model profiles are plotted in the original grid, meanwhile ensemble means are computed after linear regridding to the EN4 profile. All bias metrics in Fig. 6 use linear regridding to EN4. In Fig. 8 all data have been regridded to a 1º regular grid using nearest neighbour interpolation (closest source point). In Fig. 10 data are regridded to the ORAS5m grid, using bilinear interpolation. Figure 11 is analogous to Figs. 5 and 6, but using ORAS5m instead of EN4. In Fig. 13 data have been regridded to a 1/4º regular grid using bilinear interpolation.

- Original contribution to our understanding:

My final concern is regarding the manuscript's original contribution to our understanding of the representation of the North Atlantic Ocean in coupled climate model simulations. The authors do a good job of placing their largely qualitative findings into wider context in Section 4, however, I still remain unsure which of the study's findings are original since the impact of model resolution on sea surface biases is addressed in Roberts et al., 2019, Gutjahr et al., 2019 & Marzocchi et al., 2015, and the strength and structure of the AMOC in Talandier et al., 2014, Roberts et al., 2020, Hirschi et al., 2019, Jackson et al., 2019, Jackson et al. 2023 (see references within) and Reintges et al., 2024.

We have added new diagnostics to the manuscript (please, see next point). Nevertheless, although we acknowledge the highly valuable studies cited here, we would like to note that, compared to most previous studies, our study employs a larger, state-of-the-art mesoscale eddy-resolving model ensemble (4 models) and a larger low-resolution ensemble of CMIP6 models (39 models). This allows for a more robust characterization of the role of resolution in the representation of the North Atlantic. For example, Marzocchi et al. (2015) focus on one single, non-coupled model (NEMO). Similarly, the studies by Gutjahr et al. (2019) and Roberts et al. (2019) are based on single coupled models: MPI-ESM1.2 and HadGEM3-GC3.1, respectively. Talandier et al. (2014) employ a non-coupled ocean model (OPA), at 0.5º resolution with a nest at 0.125º resolution in the North Atlantic. Roberts et al. (2020) employ 7 coupled models but only 2 of them include mesoscale eddy-resolving configurations. Hirschi et al. (2020) use a total of 10 models, out of which, 4 are coupled, and the study focuses exclusively on AMOC analyses. The study by Reintges et al. (2024)

employs a total of ~47 different model configurations, although most are at eddy-parameterized scales: 3 of them are mesoscale eddy-permitting, but none is eddy-resolving, and that study does not investigate the effect of resolution on the representation of the AMOC. Similarly, Jackson et al. (2020)*'s study includes 5 models at eddy-parameterized and mesoscale eddy-permitting resolutions, but none at eddy-resolving resolutions.
Furthermore, the fact that we are using different models in the ensembles, compared to previous studies, constitutes an additional source of value. For example, the AMOC results obtained in our study are different from those obtained in Roberts et al. (2020) and Hirschi et al. (2020)(see lines 438-452 in the discussion of our manuscript).
Finally, we would like to note that, apart from the assessment of the surface biases and the AMOC, we have also conducted analysis on the Labrador Sea profiles and winter mixed layers in the North Atlantic. Previous work by Heuzé (2021) has investigated the mixed layer representation in a multimodel context, although that study has a different aim: the role of resolution is not assessed, and no mesoscale eddy-resolving models are considered. Koenigk et al. (2021) have successfully investigated the role of resolution on the depth of the mixed layer in a multimodel context, but they use a different definition of mixed layer and they employ 7 models, out of which one includes a mesoscale eddy-resolving configuration.
To summarize, we believe that the larger ensembles in our study, as well as the different models, methods, and combined analysis of different related quantities constitute a meaningful contribution to our common understanding of the effect of resolving the mesoscale in the North Atlantic that additionally avoids drawing model-dependent conclusions.

*We have searched for Jackson et al. (2019) but that manuscript focuses on ocean reanalysis, so we have assumed that the correct reference was Jackson et al. (2020). Please, correct us if otherwise.

To progress beyond identifying differences between ensembles and ocean observations, the study should place greater emphasis on the reasons why these differences exist, including those differences between ensemble members; for example, why is HadGEM3-GC3.1-HH often an outlier in the HR ensemble? The authors begin to address this in Section 4 by identifying an interesting 'potential link between LS salinity biases and the NAC, through the effect of the NAC on the northward salinity transport' and I would strongly encourage them to pursue this further since developing diagnostics to better understand common model biases would be a valuable contribution of this research.

We have added and discussed scatterplots (which can be found in the updated manuscript) exploring relationships between the different variables in the study, putting a special focus on salinities. 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 F6 below as an example). Besides, these new plots have been analyzed to identify and explain inter-model differences.

Specific Comments

Abstract

Lines 10: Suggest clarifying what 'standard resolution models' are? This description could be clearer for readers.

We have rephrased "which are parameterized in standard resolution models" as "which are parameterized in models with standard resolutions on the order of 1º or coarser".

Introduction

Line 39-45: Suggest revising this paragraph from one long sentence to demonstrate the interconnectivity between water mass processes. As it stands, deep convection, surface forced water mass transformation and densification along the SPG boundary are highlighted separately, yet both deep convection and boundary current densification are a result of surface forced water mass transformation. It may be beneficial to frame this discussion in terms of surface forced water mass transformation and mixing and their importance for deep convection and dense water formation – and the role of horizontal ocean model resolution in representing these processes.

The old text has been rephrased to reflect the interconnectivity between water mass processes.

New text (replacing old lines 39-45):
"The North Atlantic circulation is influenced by a series of elements and processes that are strongly interconnected. The strength and path of the NAC affect the heat and salinity content of the waters reaching the subpolar North Atlantic (SPNA; Marzocchi et al., 2015). There, the relatively warm and saline AMOC upper limb waters undergo a process of densification associated to 1) surface water mass transformation through air-sea buoyancy fluxes (Petit et al., 2020; Jackson and Petit, 2023) and 2) mixing with denser (colder) waters from the Greenland–Scotland Ridge overflows (Dey et al., 2024), together triggering deep water convection in the SPNA basins (Koenigk et al., 2021). Sinking of deep waters forming the AMOC return flow occurs at the boundaries of the SPG and it has been associated with densification of waters along the boundary current (Katsman et al., 2018; Straneo, 2006; Spall and Pickart, 2001).

The role of horizontal ocean model resolution in representing these processes is addressed in the next paragraph (from old line 46 onwards). In particular, we have added the following text to the old version:

"In the SPNA, ocean eddies contribute to the downwelling along the boundary current of the SPG through advection of density and vorticity from the interior basins (Straneo 2006; Brüggemann et al. 2017). They also contribute to restratification of Labrador Sea convective areas at the end of winter (Clément et al. 2023)."

Lines 46-55: Suggest including a brief discussion of submesoscale eddies in this paragraph and using mesoscale-resolving models to be more precise (see general comments above).

This point has been addressed in the "General Comments" section (please, see above for further details).

Lines 59-61: The Introduction appears to depend heavily on the single model study of Marzocchi et al. (2015), however, the role of ocean model resolution on the Gulf Stream position is also explored in the more recent studies of Chassignet and Xu (2017) and Chassignet et al. (2020). Suggest extending the references cited here.

We have added the reference Chassignet et al. (2020) at line 61. Regarding the discussion on the work by Chassignet and Xu (2017), we think that it fits better in the paragraph about the ocean submesoscale (see previous point and "General Comments").

Lines 78-79: Suggest rephrasing this sentence to more accurately reflect the number of multi-model comparisons that have been performed; for example, Jackson et al. (2022), Jackson et al. (2023), Reintges et al. (2024) all consider coupled climate models in a North Atlantic context.

Done. Regarding the study by Jackson et al. (2022), we have preferred not to include it in this particular list of citations, because it mostly considers forced models (only one of their figures considers coupled models). Similarly, Jackson et al. (2023) mostly focuses on the HadGEM3 model (only one figure is multimodel). The latter study will though be cited below in the first point of the "Discussion and Conclusions".

The sentence: "To our knowledge, only a few multimodel comparisons of coupled historical experiments with a focus on the North Atlantic ocean exist, that include eddy-resolving simulations (e.g. Roberts et al., 2020; Koenigk et al., 2021), although none of them specifically addresses the impact of resolving mesoscale ocean eddies.",

has been rephrased as:
"Although a wide range of multimodel studies considering coupled climate models in a North Atlantic context have been published (e.g. Reintges et al., 2024; Jackson and Petit, 2023; Bellomo et al., 2021; Heuzé, 2021; Roberts et al., 2020; Koenigk et al., 2021), only a few of them include eddy-resolving simulations (e.g. Koenigk et al., 2021; Roberts et al., 2020) and none of them specifically addresses the impact of resolving mesoscale ocean eddies."

Methods

Lines 112-119: Are the three-dimensional temperature and salinity fields used in the study stored on the original model grid or the regularly interpolated tracer fields?

This question has been addressed in the "General Comments" section (regridding details).

Here and throughout, suggest being more precise in the use of AMOC. The AMOC is a phenomenon and the overturning stream function in depth-space is a diagnostic used to understand one aspect of this phenomenon. Suggest using vertical overturning or overturning in depth-space throughout since the diapycnal overturning is not considered in this study (although I would argue is a more relevant diagnostic to consider due to its close relationship to sea surface property biases explored later in this study).

The term "AMOC" has been replaced with the terms "overturning in depth-space" or "AMOC streamfunction" throughout the manuscript.

Lines 141-144: The authors make a strong case for diagnosing the MLD from the time-averaged potential density anomaly field following de Boyer Montegut et al. (2004), so I was surprised that the authors did not then compare this result to the available de Boyer Montegut et al. (2004) MLD climatology.

We aimed at producing MLD estimates both from models and observations that 1) were calculated using the same methodology and that 2) were based on data with similar characteristics, i.e., gridded data. That is why we made the choice of using EN4 gridded data instead of MLD observational products based on individual profiles. Now ARMOR3D MLDs calculated with the same methodology as in models (from temperature and salinity data and a density threshold of 0.03 kg m^{-3}) have also been added to the MLD analysis to provide an additional comparison (please, refer to Fig. F1)
However, qualitative comparisons to observational MLD data are carried out in the discussion section of our manuscript (lines 410-418).

Lines 152-156: How is the AMOC & barotropic stream function calculated in the ORAS5m reanalysis data? Is the calculation performed on the original model grid or using interpolated model fields?

Both the AMOC and the barotropic stream function in the ORAS5m data were calculated on the original model grid. We have added this information in the text.

Results

Figures 2-4: Here and throughout the manuscript text, suggest discussing the statistical significance of the differences between the LR & HR ensembles. For example, there is considerable discrepancies between SST & SSS bias within the HR ensemble, especially around the NAC. Is the improvement in SST bias in the Central North Atlantic region in the HR ensemble simply due to the warm bias exhibited by HadGEM3-GC31-HH counteracting the cold biases in the other ensemble members?

To address the statistical significance of the differences in means between the LR & HR ensembles, 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).

As a first analysis, bootstrapping has been applied using maximum ensemble sizes in both the LR and the HR samples, i.e. by taking LR ensemble samples with the total size of the LR ensemble, and HR ensemble samples with the total size of the HR ensemble (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, a second 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). This second analysis is aimed at investigating whether the differences in means between the LR & HR ensembles are still significant when the LR ensemble is considered as subsamples of the same size as the HR ensemble (which is considerably smaller than the LR ensemble), i.e. whether by randomly picking subsamples from the LR ensemble of the same size as the HR ensemble (usually 4), the differences in means between LR & HR are still significant.

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. This allowed us to determine the minimum size of the LR ensemble required for significance (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).

As pointed out by the reviewer, 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). In order to test whether the reduction in the CNA SST bias in HR-HIST is related to the warm bias shown by HadGEM3-GC31-HH in that region, we have removed the HadGEM3-GC31-HH model from the HR ensemble. In that case, the SST HR mean in the CNA is still larger than the LR mean, but the difference is very small (LR-HIST SST mean = -2.83 ºC vs HR-HIST mean = -2.47 ºC) and again not significant at the 95% confidence level. 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).

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 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.

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.

Lines 255-256: Is the correlation between MLD (deep convection) and AMOC (assuming you are referring to vertical overturning strength) in models in Martin-Martinez et al.? This relationship is much less clear in observations (see Li et al., 2021 for discussion in relation to the OSNAP observing system).

Yes, Martin-Martinez et al. (2024) show a correlation between March MLDs in the Labrador/Irminger Sea and the AMOC streamfunction at different time lags and latitudes for the

EC-Earth-3P model across different resolutions (Fig. 7 in that manuscript; link to preprint: https://egusphere.copernicus.org/preprints/2024/egusphere-2024-3625/egusphere-2024-3625.pdf). Results from a multimodel study (Li et al., 2019) also show correlations between March Labrador Sea MLDs and AMOC strength at different latitudes (Fig. 11 in that study), as well as between winter-averaged Labrador Sea Water density and AMOC at different latitudes (Fig. 9 in that study). Similarly, Menary et al. (2020) show a lagged correlation between SPNA densities (between 500-1500 m depth) and AMOC strength at 45ºN (Fig. 3a in that manuscript) in the HadGEM3-GC3.1-MM model.
Li et al. (2021)'s study supports a relationship between density anomalies in the SPNA and AMOC variability, which we believe does not contradict our original statement.
We have extended the references in the old text to include that of Li et al. (2019) and added a remark on the interesting results by Li et al. (2021).

New text:
"MLD is generally used as a proxy for deep water convection, it is correlated to AMOC strength (Martin-Martinez et al., in review; Li et al., 2019), and in the North Atlantic it achieves its maximum in March. We note that results from a recent observational study suggest that AMOC variability depends on combined density anomalies from different areas of the SPNA rather than from one location alone (Li et al., 2021), such that MLD analyses in the SPNA should not be limited to one single region (e.g. the central Labrador Sea).

Old text:
"MLD is generally used as a proxy for deep water convection, it is correlated to AMOC strength (Martin-Martinez et al., in review), and in the North Atlantic achieves its maximum in March"

Lines 264-265: Given that EN4.2.2 is too coarse & has insufficient observational data (Argo etc.) to resolve subpolar boundary currents, can you really assess if the convection region along the Irminger Sea western boundary current is better represented in the HR ensemble?

To improve our picture of the representation of mixed layers in the Irminger Sea and in the SPNA in general, we have added the finer resolution ARMOR3D dataset to our analysis (please, see the "General Comments" section). We agree with the reviewer that the 1º EN4 dataset cannot resolve all the details of the East Greenland current, but it can characterize it broadly. We have tested the sensitivity of the EN4 MLDs to the observational period (Argo vs pre-Argo) (Fig. F4) and the convection patterns are similar in both cases. However, we observe that the western IS and the LS convection areas are less connected for the 2000-2014 (Argo) period compared to the 1980-1999 (pre-Argo) period in EN4, which could be related to natural variability but also to the more restricted availability of EN4 profiles prior the year 1999. We have edited our old text accordingly (see below).
Note: this new figure will be added to the manuscript as a supplementary figure.

New text:
"Notice too, that the convection area along the East Greenland current, in the western IS, is also deeper in the HR-HIST ensemble with respect to LR-HIST, better resembling the EN4 pattern."

Old text:
"Notice too, that the convection area along the East Greenland current, in the western IS, is also deeper and better connected with the mixed layers of the LS in the HR-HIST ensemble with respect to LR-HIST, better resembling the EN4 pattern."

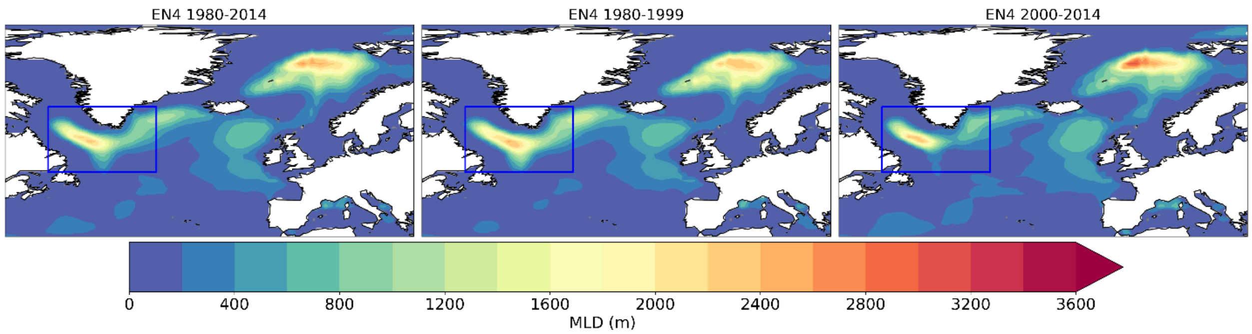

Figure F4. EN4 March MLD (in m) for the (a) 1980-2014, (b) 1980-1999 (pre-Argo) and (c) 2000-2014 periods.

Figure 9: Is the AMOC vertical overturning stream function in the HR ensemble statistically significantly different from the LR ensemble, given the wide range of AMOC mean states shown in Figure 9 (LR ensemble panels).

This remark has been addressed in a previous point (see bootstrapping analysis above).

Lines 293-294: When comparing model results to the RAPID-MOCHA array is the 2004-2022 period used or the 2004-2014 period overlapping the end of the historical simulations?

In Fig. 11 we compared the 1980-2014 period in models and ORAS5m reanalysis with the 2004-2022 period in RAPID. We have now produced an additional plot to test the sensitivity of the results to the choice of the RAPID period (Fig. F5). In this new figure the RAPID and GLORYS12 data cover the intervals 2004-2014 and 1993-2014, respectively, overlapping the end of the historical simulations. Figure F5 will replace old Fig. 11 in the manuscript. However, differences with respect to the original analysis do not appear significant.

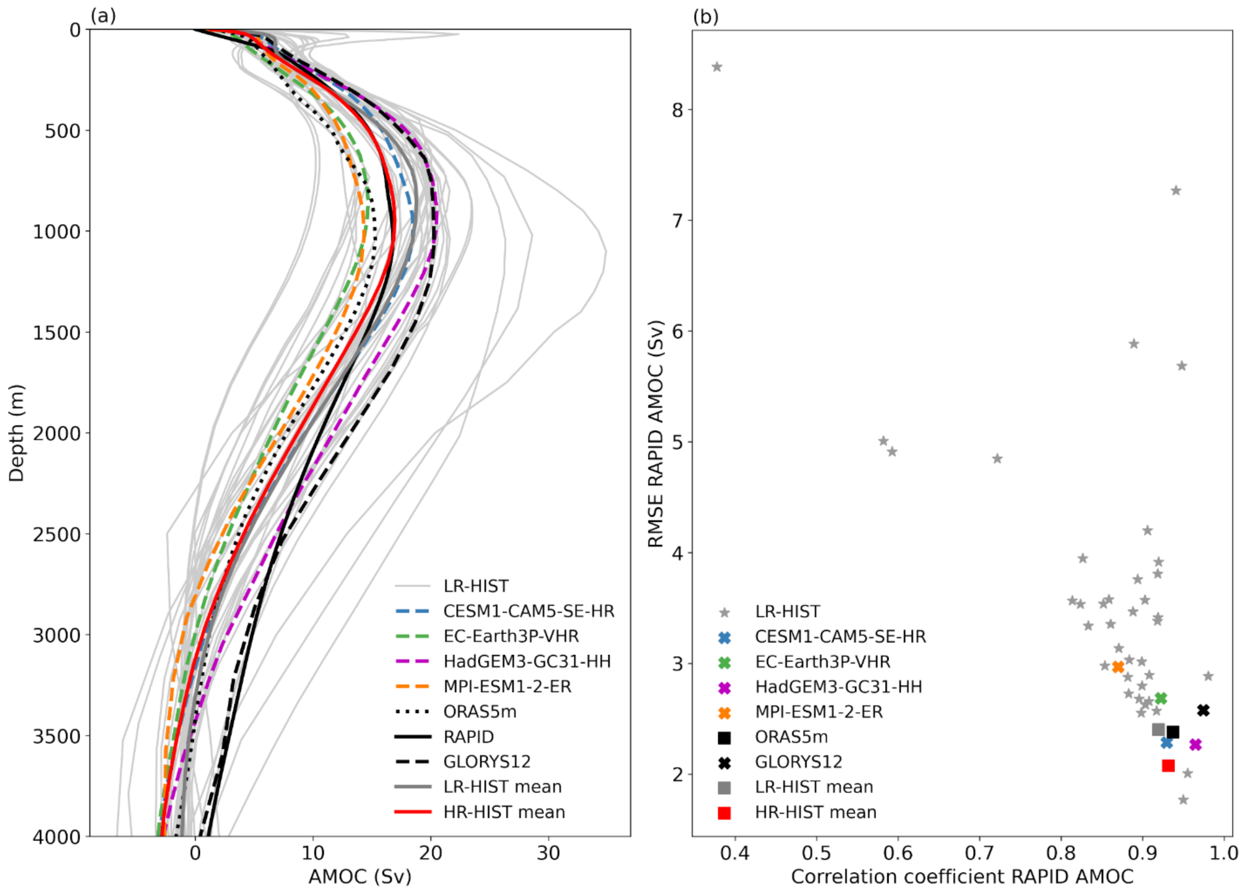

Figure F5. (a) Climatological AMOC profile at 26.5º N (in Sv). (b) Pearson correlation coefficient (horizontal axis) and Root Mean Square Error (RMSE; in Sv) (vertical axis) of AMOC profiles at 26.5º N against RAPID, both estimated across the vertical dimension. In both subplots, models and ORAS5m reanalysis data correspond to the interval 1980–2014, RAPID observations are averaged over the period April 2004 – December 2014, and GLORYS12 reanalysis over the interval 1993-2014.

Lines 305-306: Suggest being more specific on the differences in the methodological approach; was the RAPID overturning stream function calculated using the METRIC package or are you comparing the model 'truth' to the RAPID calculation applied to observations?

We compare the model "truth" to RAPID calculations. More details on the methodological approach have been added to the text. Please, see below.

New text (replacing lines 296-306):

" 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 streamfunction 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 (lines 296-306):
"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)."

Lines 328-344: When discussing the Gulf Stream path, the predominant focus is on the location of separation with only limited commentary on the current's structure. Suggest undertaking a more detailed evaluation of the Gulf Stream structure, including exploring its eastward penetration using surface eddy kinetic energy following Xu and Chassignet (2017). Alternatively, an observational product such as COPERNICUS-GLOBCURRENT could be used to validate the model surface current velocities.

Thanks for this remark. We would prefer performing new diagnostics based on the already available/processed variables, because processing an additional variable (such as eddy kinetic energy) for the 44 models analyzed in this study is an arduous task due to ESMValTool constraints, which requires extremely strict cmorization of the data. We believe however, that this analysis is not indispensable for the completeness of our study.
Additionally, we would like to note that in the original manuscript we have analyzed the location of separation of the Gulf Stream from the coast but also the NAC structure through the barotropic streamfunction (Figs. 12 and 13). We have validated the position of the NAC using absolute dynamic topography (old lines 318-321). Besides, in (old) lines 345-350 we have associated a less zonal NAC structure at high resolution with a reduced bias in the CNA region.

Lines 353-356: This is an interesting point, suggest exploring the relationship between interior dense water formation and SPG dynamics further as a potential explanation for ensemble spread / differences.

We have added a scatterplot showing the correlation between MLD and SPG strength in the different models (Fig. F6), and discussed the results (the discussion is included in the updated manuscript).

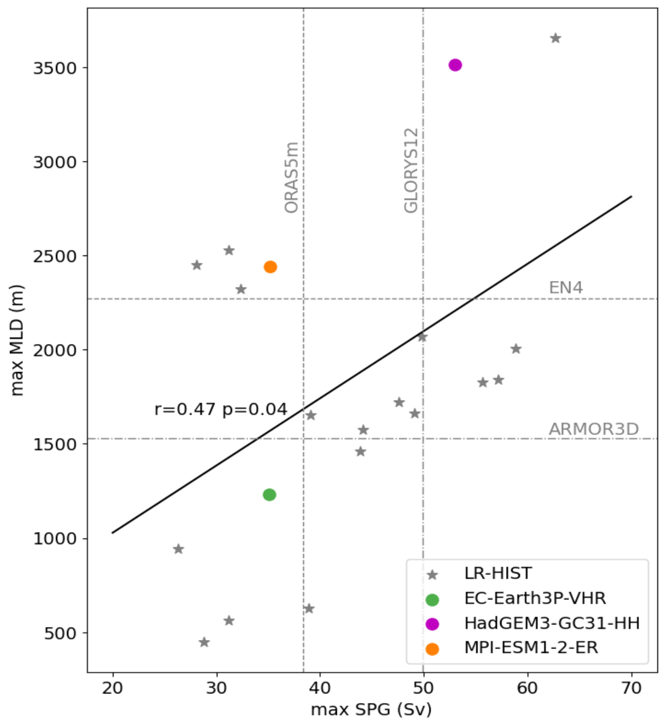

Figure F6. Scatterplot of max. SPG strength (in Sv) vs max. MLD (in m), both referred to the LIS box (shown in Fig. 8). Pearson correlation coefficient and p-value are shown next to the fit line. Horizontal dashed and dot-dashed lines show EN4 and ARMOR3D observation-based values, respectively. Vertical dashed and dot-dashed lines show ORAS5m and GLORYS12 reanalysis values, respectively.

Discussion and Conclusions

Lines 388-391: The current discussion of Labrador Sea biases should be revised to cite more recent perspectives (e.g., review by Jackson et al., 2023; Li et al., 2023; Rühs et al. 2021).

The text has been modified to reflect the findings by the above-mentioned studies (see new discussion in the updated manuscript).

Lines 391-394: This is an interesting hypothesis to link the NAC and LS salinity biases. Suggest reading Kostov et al. (2023, 2024) on the connection between the NAC and LS convection to extend these ideas further in the manuscript as suggested in General Comments.

A discussion has been included on the results by Kostov et al. (2023, 2024) in relation to our findings (see updated manuscript).

Lines 408-412: Given the limitations you have identified with the EN4.2.2 product, why not use a 'purpose-built' global mixed layer climatology, such as the LOPS-IFREMER MLD product (https://doi.org/10.17882/98226)? Note, this is still a coarse resolution product, so using ARMOR3D may be more appropriate to compare to HR models.

Done, the ARMOR3D MLDs have been added to the analysis (please, refer to the "General Comments" section).

Lines 467-470: Suggest revising this summary to focus on why the HR ensemble shows improvements compared to ocean observations, and highlight next steps forward in coupled climate modelling; for example, what are the implications of this improved representation of the ocean mean state in mesoscale-resolving ocean models on the atmosphere and societally relevant indicators? This invokes a wider question of whether the improvements in the North Atlantic mean state are sufficient to justify the additional computational cost, and to what extent the mean state determines the ocean's future trajectory in coupled models.

Done, the text has been modified to add a discussion of all these points (see updated manuscript).

References

Bellomo, K., Angeloni, M., Corti, S., and von Hardenberg, J.: Future climate change shaped by inter-model differences in Atlantic meridional overturning circulation response, Nat Commun, 12, 3659, https://doi.org/10.1038/s41467-021-24015-w, 2021.

de Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A., and Iudicone, D.: Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology, J. Geophys. Res. Oceans, 109, C12003, https://doi.org/10.1029/2004JC002378, 2004.

Bruciaferri, D., Guiavarc'h, C., Hewitt, H. T., Harle, J., Almansi, M., Mathiot, P., and Colombo, P.: Localized General Vertical Coordinates for Quasi-Eulerian Ocean Models: The Nordic Overflows Test-Case, Journal of Advances in Modeling Earth Systems, 16, e2023MS003893, https://doi.org/10.1029/2023MS003893, 2024.

Brüggemann, N., Katsman, C. A., and Dijkstra, H. A.: On the vorticity dynamics of the downwelling branch of the AMOC, CLIVAR Exchanges Special Issue: CLIVAR Open Science Conference Award Winners, 71, 10-12, 2017.

Chassignet, E. P. and Xu, X.: Impact of Horizontal Resolution (1/12° to 1/50°) on Gulf Stream Separation, Penetration, and Variability, https://doi.org/10.1175/JPO-D-17-0031.1, 2017.

Chassignet, E. P., Yeager, S. G., Fox-Kemper, B., Bozec, A., Castruccio, F., Danabasoglu, G., Horvat, C., Kim, W. M., Koldunov, N., Li, Y., Lin, P., Liu, H., Sein, D. V., Sidorenko, D., Wang, Q., and Xu, X.: Impact of horizontal resolution on global ocean–sea ice model simulations based on the experimental protocols of the Ocean Model Intercomparison Project phase 2 (OMIP-2), Geoscientific Model Development, 13, 4595–4637, https://doi.org/10.5194/gmd-13-4595-2020, 2020.

Clément, L., Frajka-Williams, E., Oppeln-Bronikowski, N. von, Goszczko, I., and Young, B. de: Cessation of Labrador Sea Convection Triggered by Distinct Fresh and Warm (Sub)Mesoscale Flows, https://doi.org/10.1175/JPO-D-22-0178.1, 2023.

Colombo, P., Barnier, B., Penduff, T., Chanut, J., Deshayes, J., Molines, J.-M., Le Sommer, J., Verezemskaya, P., Gulev, S., and Treguier, A.-M.: Representation of the Denmark Strait overflow in a *z*-coordinate eddying configuration of the NEMO (v3.6) ocean model: resolution and parameter impacts, Geoscientific Model Development, 13, 3347–3371, https://doi.org/10.5194/gmd-13-3347-2020, 2020.

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.

Dey, D., Marsh, R., Drijfhout, S., Josey, S. A., Sinha, B., Grist, J., and Döös, K.: Formation of the Atlantic Meridional Overturning Circulation lower limb is critically dependent on Atlantic-Arctic mixing, Nat Commun, 15, 7341, https://doi.org/10.1038/s41467-024-51777-w, 2024.

Guinehut, S., Dhomps, A.-L., Larnicol, G., and Le Traon, P.-Y.: High resolution 3-D temperature and salinity fields derived from in situ and satellite observations, Ocean Science, 8, 845–857, https://doi.org/10.5194/os-8-845-2012, 2012.

Gutjahr, O., Putrasahan, D., Lohmann, K., Jungclaus, J. H., von Storch, J.-S., Brüggemann, N., Haak, H., and Stössel, A.: Max Planck Institute Earth System Model (MPI-ESM1.2) for the High-Resolution Model Intercomparison Project (HighResMIP), Geosci. Model Dev., 12, 3241–3281, https://doi.org/10.5194/gmd-12-3241-2019, 2019.

Heuzé, C.: Antarctic Bottom Water and North Atlantic Deep Water in CMIP6 models, Ocean Science, 17, 59–90, https://doi.org/10.5194/os-17-59-2021, 2021.

Hirschi, J. J.-M., Barnier, B., Böning, C., Biastoch, A., Blaker, A. T., Coward, A., Danilov, S., Drijfhout, S., Getzlaff, K., Griffies, S. M., Hasumi, H., Hewitt, H., Iovino, D., Kawasaki, T., Kiss, A. E., Koldunov, N., Marzocchi, A., Mecking, J. V., Moat, B., Molines, J.-M., Myers, P. G., Penduff, T., Roberts, M., Treguier, A.-M., Sein, D. V., Sidorenko, D., Small, J., Spence, P., Thompson, L., Weijer, W., and Xu, X.: The Atlantic Meridional Overturning Circulation in High-Resolution Models, Journal of Geophysical Research: Oceans, 125, e2019JC015522, https://doi.org/10.1029/2019JC015522, 2020.

Jackson, L. C. and Petit, T.: North Atlantic overturning and water mass transformation in CMIP6 models, Clim Dyn, 60, 2871–2891, https://doi.org/10.1007/s00382-022-06448-1, 2023.

Jackson, L. C., Roberts, M. J., Hewitt, H. T., Iovino, D., Koenigk, T., Meccia, V. L., Roberts, C. D., Ruprich-Robert, Y., and Wood, R. A.: Impact of ocean resolution and mean state on the rate of AMOC weakening, Clim Dyn, 55, 1711–1732, https://doi.org/10.1007/s00382-020-05345-9, 2020.

Jackson, L. C., Biastoch, A., Buckley, M. W., Desbruyères, D. G., Frajka-Williams, E., Moat, B., and Robson, J.: The evolution of the North Atlantic Meridional Overturning Circulation since 1980, Nat Rev Earth Environ, 3, 241–254, https://doi.org/10.1038/s43017-022-00263-2, 2022.

Jackson, L. C., Hewitt, H. T., Bruciaferri, D., Calvert, D., Graham, T., Guiavarc'h, C., Menary, M. B., New, A. L., Roberts, M., and Storkey, D.: Challenges simulating the AMOC in climate models, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 381, 20220187, https://doi.org/10.1098/rsta.2022.0187, 2023.

Katsman, C. A., Drijfhout, S. S., Dijkstra, H. A., and Spall, M. A.: Sinking of Dense North Atlantic Waters in a Global Ocean Model: Location and Controls, Journal of Geophysical Research: Oceans, 123, 3563–3576, https://doi.org/10.1029/2017JC013329, 2018.

Koenigk, T., Fuentes-Franco, R., Meccia, V. L., Gutjahr, O., Jackson, L. C., New, A. L., Ortega, P., Roberts, C. D., Roberts, M. J., Arsouze, T., Iovino, D., Moine, M.-P., and Sein, D. V.: Deep mixed ocean volume in the Labrador Sea in HighResMIP models, Clim Dyn, 57, 1895–1918, https://doi.org/10.1007/s00382-021-05785-x, 2021.

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.

Li, F., Lozier, M. S., Danabasoglu, G., Holliday, N. P., Kwon, Y.-O., Romanou, A., Yeager, S. G., and Zhang, R.: Local and Downstream Relationships between Labrador Sea Water Volume and North Atlantic Meridional Overturning Circulation Variability, https://doi.org/10.1175/JCLI-D-18-0735.1, 2019.

Li, F., Lozier, M. S., Bacon, S., Bower, A. S., Cunningham, S. A., de Jong, M. F., deYoung, B., Fraser, N., Fried, N., Han, G., Holliday, N. P., Holte, J., Houpert, L., Inall, M. E., Johns, W. E., Jones, S., Johnson, C., Karstensen, J., Le Bras, I. A., Lherminier, P., Lin, X., Mercier, H., Oltmanns, M., Pacini, A., Petit, T., Pickart, R. S., Rayner, D., Straneo, F., Thierry, V., Visbeck, M., Yashayaev, I., and Zhou, C.: Subpolar North Atlantic western boundary density anomalies and the Meridional Overturning Circulation, Nat Commun, 12, 3002, https://doi.org/10.1038/s41467-021-23350-2, 2021.

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.

Menary, M. B., Jackson, L. C., and Lozier, M. S.: Reconciling the Relationship Between the AMOC and Labrador Sea in OSNAP Observations and Climate Models, Geophysical Research Letters, 47, e2020GL089793, https://doi.org/10.1029/2020GL089793, 2020.

Moreno-Chamarro, E., Arsouze, T., Acosta, M., Bretonnière, P.-A., Castrillo, M., Ferrer, E., Frigola, A., Kuznetsova, D., Martin-Martinez, E., Ortega, P., and Palomas, S.: The very-high-resolution configuration of the EC-Earth global model for HighResMIP, Geoscientific Model Development, 18, 461–482, https://doi.org/10.5194/gmd-18-461-2025, 2025.

Petit, T., Lozier, M. S., Josey, S. A., and Cunningham, S. A.: Atlantic Deep Water Formation Occurs Primarily in the Iceland Basin and Irminger Sea by Local Buoyancy Forcing, Geophysical Research Letters, 47, e2020GL091028, https://doi.org/10.1029/2020GL091028, 2020.

Reintges, A., Robson, J. I., Sutton, R., and Yeager, S. G.: Subpolar North Atlantic Mean State Affects the Response of the Atlantic Meridional Overturning Circulation to the North Atlantic Oscillation in CMIP6 Models, https://doi.org/10.1175/JCLI-D-23-0470.1, 2024.

Roberts, C. D., Garry, F. K., and Jackson, L. C.: A Multimodel Study of Sea Surface Temperature and Subsurface Density Fingerprints of the Atlantic Meridional Overturning Circulation, https://doi.org/10.1175/JCLI-D-12-00762.1, 2013.

Roberts, M. J., Baker, A., Blockley, E. W., Calvert, D., Coward, A., Hewitt, H. T., Jackson, L. C., Kuhlbrodt, T., Mathiot, P., Roberts, C. D., Schiemann, R., Seddon, J., Vannière, B., and Vidale, P. L.: Description of the resolution hierarchy of the global coupled HadGEM3-GC3.1 model as used in CMIP6 HighResMIP experiments, Geosci. Model Dev., 12, 4999–5028, https://doi.org/10.5194/gmd-12-4999-2019, 2019.

Roberts, M. J., Jackson, L. C., Roberts, C. D., Meccia, V., Docquier, D., Koenigk, T., Ortega, P., Moreno-Chamarro, E., Bellucci, A., Coward, A., Drijfhout, S., Exarchou, E., Gutjahr, O., Hewitt, H., Iovino, D., Lohmann, K., Putrasahan, D., Schiemann, R., Seddon, J., Terray, L., Xu, X., Zhang, Q., Chang, P., Yeager, S. G., Castruccio, F. S., Zhang, S., and Wu, L.: Sensitivity of the Atlantic Meridional Overturning Circulation to Model Resolution in CMIP6 HighResMIP Simulations and Implications for Future Changes, Journal of Advances in Modeling Earth Systems, 12, e2019MS002014, https://doi.org/10.1029/2019MS002014, 2020.

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.

Spall, M. A. and Pickart, R. S.: Where Does Dense Water Sink? A Subpolar Gyre Example, 2001.

Straneo, F.: On the Connection between Dense Water Formation, Overturning, and Poleward Heat Transport in a Convective Basin, https://doi.org/10.1175/JPO2932.1, 2006.

Tagklis, F., Bracco, A., Ito, T., and Castelao, R. M.: Submesoscale modulation of deep water formation in the Labrador Sea, Sci Rep, 10, 17489, https://doi.org/10.1038/s41598-020-74345-w, 2020.

Talandier, C., Deshayes, J., Treguier, A.-M., Capet, X., Benshila, R., Debreu, L., Dussin, R., Molines, J.-M., and Madec, G.: Improvements of simulated Western North Atlantic current system and impacts on the AMOC, Ocean Modelling, 76, 1–19, https://doi.org/10.1016/j.ocemod.2013.12.007, 2014.