The authors sincerely acknowledge both reviewers for their comments, as well as for the time and effort spent in revising our responses.

Responses are in blue and changes in the manuscript in green.

Responses to referee#1

Summary

I would like to thank the authors for having taken all my comments into account and for their considerable efforts in revising the manuscript. This revised version of the manuscript is much improved, especially through the additions of the new observational analysis & reanalysis products, and the more robust statistical assessment of the LR & HR ensembles.

The authors have resolved all my original concerns. Upon addressing my small number of minor comments below, I would recommend this manuscript be accepted for publication in Ocean Sciences.

General Comments

§ Manuscript Length: Whilst I'm grateful to the authors for addressing the comments of Reviewer #2 and myself so thoroughly, I'm slightly concerned that the revised manuscript has increased significantly in length from 34 to 47 total pages. I would strongly recommend reviewing the manuscript for opportunities to consolidate the existing figures / text, whilst preserving the central messages.

The length of the manuscript has been shortened. The following changes have been made to shorten the text:

1- The discussion about MLD values in the LS (old lines 339-347 + old lines 698-722) has now been shortened (now lines 346-367).

2- Table 2 has been moved to Appendix A and renamed as Table A1.

3- Fig. 16 has been moved to Appendix A (as it was in the original draft, before the revisions) and renamed as Fig. A3. Old Fig. A3 is now Fig. A4.

4- The discussion about the RAPID method (old lines 392-413) has been shortened.

New text replacing old lines 392-413 (now at new lines 424-431):

"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 at every grid point, this approach is not possible with observations, since direct velocity measurements are scarce. The calculation of the upper mid-ocean return transport in RAPID is based on the zonal gradient of dynamic heights from density profiles, which makes use of a reference depth (4820 m), representing a level-of-no-motion (Roberts et al., 2013; McCarthy et al., 2015; Danabasoglu et al., 2021). 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)."

5- Old lines 775-780 have been removed following a suggestion by another reviewer.

I'd like to highlight one possible approach to do this below (although this is simply a suggestion):

o Distributing the findings of 3.6 Testing the significance of differences between ensembles amongst the relevant sections earlier in the Results. As a reader, I felt that this statistical analysis would have been most useful when the diagnostics were first discussed. This would provide an opportunity to reduce the length of the text, and the summary table could be moved to Supplementary Information / Appendix A.

The summary table has been moved to Appendix A following the reviewer's suggestion. Regarding redistributing section 3.6 into the previous sections, we believe it would probably not contribute to shorten the length of the manuscript, since, for example, in section 3.6, the significance of changes in the SST and SSS biases between ensembles are treated jointly (and thus occupy less space), meanwhile in the previous analyses they are split between sect. 3.1 and sect. 3.2.

o Similarly, the contents of 3.7 Characteristic features in the HR-HIST models could be distributed between the previous Results sections (for quantitative comparisons – which would then be helpfully located near each large multi-panel figure) and the excellent Discussion & Conclusions section (for the more speculative discussion points).

We believe it is useful to have a section where the different metrics of a specific model are discussed altogether, it provides the opportunity to relate them to each other (which is difficult in separate sections analysing separate metrics). For this reason, we would prefer to keep sect. 3.7 as a whole.

Importantly, we would also like to note that the length of the manuscript reflects a comprehensive analysis, including several multimodel figures and mesoscale-resolving models.

§ 3.8 Relations between dynamical and physical properties:

This is an interesting addition to the manuscript, which directly addresses my previous concerns regarding the originality of the study. My only concern is whether it is appropriate to use a composite of the LR-HIST and HR-HIST ensembles to explore the correlations between state variables, since earlier you highlight important differences between these ensembles. For example, might it be the case that the relationship between the SPG strength and the maximum overturning at 26.5N is different between the LR & HR ensembles? Hirschi et al. (2020) highlighted that, at HR, the more realistic SPG circulation projects more strongly onto the diapycnal rather than vertical overturning at subpolar latitudes (horizontal circulation across sloping isopycnals), whereas LR models exhibit a more classical vertical "conveyor" like overturning cell.

Suggest commenting that the relationships between dynamical & physical properties may hence also depend on horizontal resolution (although a larger HR ensemble would be needed to perform this analysis). This may be a case of bringing some of the excellent discussion on Lines 765-774 forward in the text.

We have added the following paragraph at former line 612 (now line 663):

"Horizontal resolution might play a role in the representation of the relationships between the different dynamical and physical properties in the North Atlantic through differences in model dynamics. For example, work by Katsman et al. (2018) shows differences in deep water sinking mechanisms at mesoscale-permitting resolutions (see Sect. 4 for further details), which might affect relationships involving overturning. In this study we cannot properly assess whether such relationships change with resolution given the limited size of the HR-HIST ensemble, but future studies might be able to address it as new mesoscale-resolving simulations become available".

Specific Comments

Methods

Use of GLORYS12v1: I'm grateful to the authors for including the mesoscale eddy resolving GLORYS12v1 reanalysis product in their model evaluation, however, I have some concerns regarding the implications of using the regridded outputs for the purpose of calculating meridional overturning and barotropic stream functions. Given that GLORYS12v1 is originally simulated on a curvilinear ORCA12 grid, the use of linear interpolation to regrid the model velocity will inevitably introduce multiple sources of error into a volume transport calculation (including estimation of grid cell areas, interpolation errors at high-latitudes and considerations of bottom topography). This likely also applies to the ORAS5m outputs shown. Given that GLORYS12v1 is available on its original NEMO model grid, it would at least be worth briefly

emphasising to readers the limitations of using a regridded velocity field, although I suspect this will not alter your qualitative results.

We have now commented on the limitations of the GLORYS12v1 overturning and barotropic streamfunction data (at old line 202; now line 207). Regarding ORAS5m data, these were available in the original grid, which was the one used in our calculations, and thus those data do not present the same limitations as our GLORYS12v1 dataset.

New text (at new line 207):

"As a second reference, in addition to ORAS5m, GLORYS12 reanalysis data at mesoscale eddy-resolving resolution (1/12º; period 1993–2014; Lellouche et al., 2018) are also employed to add robustness to our analyses. In the case of ORAS5m, the overturning and barotropic streamfunctions are calculated from velocity fields in the original reanalysis ocean model grid, while for GLORYS12, they are calculated based on the regridded velocity fields available from Copernicus. We note that the use of regridded fields for volume transport calculations might introduce some errors related to, for example, the estimates of grid cell areas. However, the GLORYS12 data still constitute a valuable qualitative reanalysis reference."

Old text:

"As a second reference, in addition to ORAS5m, GLORYS12 reanalysis data at mesoscale eddy-resolving resolution (1/12º; period 1993–2014; Lellouche et al., 2018) are also employed to add robustness to our analyses. In the case of ORAS5m, the overturning and barotropic streamfunctions are calculated from velocity fields in the original reanalysis ocean model grid, while for GLORYS12, they are calculated based on the regridded velocity fields available from Copernicus."

Results

Figure 5: Suggest using a different colour / marker to identify the LR-HIST mean profile than dark grey, this is quite difficult to distinguish from the large number of LR-HIST ensemble member profiles (Fig. 5b especially) - although I do recognise the need to relate these visually.

We would prefer to keep the original colour scheme, which yields the best compromise we could achieve to provide visibility to the HR models against the benchmark of LR models (and still relate them visually). The final jpg figure will have a much improved resolution than it has now in the submitted pdf. The fact that the different profiles might not be so easy to distinguish from one another in the deep ocean has to do with the fact that there is significant overlapping at depth, but this would also be the case with a different colour scheme. More importantly, note that

differences between the HR and the LR ensemble means occur in the upper ocean, and those are clear in the current figure.

Lines 392-419: Suggest revising this discussion on the methodological concerns regarding comparisons between models and observations at the RAPID array. I would highlight two important points that are missing:

1. There is an existing approach to compare ocean models with RAPID observations in an equivalent manner, accounting for the four separate AMOC components: using the Meridional ovErTurning ciRculation diagnostIC (METRIC) package originally developed by Castrucio et al. (https://github.com/AMOCcommunity/metric).

2. The chosen treatment of the net throughflow across the RAPID 26.5N section can be a important control on the magnitude of the depth-space overturning calculated in models. It would be useful to inform readers whether a uniform volume transport compensation term was applied to each of the models (as done in observations) prior to the stream function calculation or if these diagnostics include the net throughflow across the section. This would be especially relevant for the reanalysis data used, since the regridded velocity field does not properly represent bathymetry.

We have added the following text at old line 417 (now at lines 457-462):

"Model and reanalysis overturning profiles at 26.5º N in our study are computed from full velocities and do not include the application of a uniform volume transport compensation term. This might be relevant for the GLORYS12 overturning profile, which is based on regridded velocities (see Sect. 2.3). We would like to note the existence of the software package Meridional ovErTurning ciRculation diagnostIC (METRIC), for calculating RAPID observations-equivalent AMOC diagnostics in models using different model output variables (Castruccio, 2021; Danabasoglu et al., 2021)".

New citation in references:

Castruccio, F. S.: NCAR/metric: metric v0.1., Zenodo [code], https://doi.org/10.5281/zenodo.4708277, 2021.

We have also moved old lines 417-419 to new lines 422-424.

Responses to referee#2

Thank you for the revised manuscript "The North Atlantic mean state in mesoscale eddy-resolving coupled models: a multimodel study". The paper is well written and much improved on the previous version. Though there are no major new findings, it is a useful reference for impacts of resolution, particularly for mesoscale eddy-resolving, which hasn't been covered by many studies. I have a few minor comments.

L53 'it has' -> have

Here 'it' refers to the 'sinking', which is singular, not to the 'deep waters'. For this reason, we would prefer to keep the singular form of the verb.

New text (at new line 52):

"Sinking of deep waters forming the AMOC return flow occurs at the boundaries of the subpolar gyre (SPG) and has been associated with densification of waters along the boundary current (Katsman et al., 2018; Straneo, 2006; Spall and Pickart, 2001)."

Old text:

"Sinking of deep waters forming the AMOC return flow occurs at the boundaries of the subpolar gyre (SPG) and it has been associated with densification of waters along the boundary current (Katsman et al., 2018; Straneo, 2006; Spall and Pickart, 2001)."

L279-288 Do models also generally have the same SST and SSS biases in regions outside the LS?

In the regions outside the LS, i.e. in the CNA and NCH regions, models present biases of the same sign, positive in the NCH region and negative in the CNA region. In this sense it is not possible to separate models according to the sign of their biases.

L340 'in consistency' -> consistent

Changed (now line 347).

L341-345 I think the discussion about MLD datasets in the discussion should be moved here. At the moment you start by saying that HR MLD is in better agreement with EN4 than LR. I think

you need to say that the observations of MLD vary a lot (probably because of differences in methodology – this could probably be shortened). Then either say that you can't assess because of the wide range of observations, or make an argument than EN4 is most like the models (because you're using the same methodology) and then compare to the models.

The discussion about the MLD datasets in Section 4 has been moved here. Besides, details about the different methodologies have been shortened. The conclusions that we have been able to extract are also described in the new text below.

New text in the results section (replacing old lines 339-347; now at lines 346-367):

"In the LS, the multi-model mean of HR-HIST shows a deeper (although not significantly deeper; see Sect. 3.6) mixed layer than the LR-HIST mean (Fig. 8), consistent with a relatively weaker density stratification (Sect. 3.2). If we check the individual models (Fig. 7) we note that all the HR-HIST models show deep mixed layers in the LS, while ~25 % of the LR-HIST models show little or no convection. LS mixed layers in the HR-HIST mean (1800–2000 m) are closer to EN4 estimates (2000–2200 m), whereas in ARMOR3D (1000–1200 m) they are in the same range as in the LR-HIST mean (1000–1200 m). For the overlapping time interval of ARMOR3D and EN4 (i.e., for 1993–2014), EN4 values for the LS are still larger (1800–2000 m; not shown) compared to ARMOR3D. The wide range in the observation-derived estimates for the LS in our analyses leads us to review the literature for observational studies. Work by Holte et al. (2017) based on individual Argo density profiles shows mixed layers down to 1400–1800 m in the LS for the 2000–2016 period. Time-varying estimates of winter maximum MLDs in the LS obtained from Argo floats, the AR7W line, and moored measurements, suggest values mostly around 1100–1500 m in the 2002–2015 interval (Yashayaev and Loder, 2016), showing an intensification in recent years, with a record value of 2100 m in 2016 (Yashayaev and Loder, 2017). Therefore, the MLD values in our HR-HIST (1800–2000 m) ensemble mean are slightly too large compared to observational studies, and the LR-HIST (1000–1200 m) ensemble mean values are slightly too shallow. The differences in the MLD estimates obtained across the different studies might arise from the different temporal and spatial characteristics of the profile data, differences in the time intervals analysed, and from the different methodological approaches employed. 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. 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 a combination of methods and elements (including temperature and density threshold methods, gradient methods, estimates of thermocline linear fits, etc). In our study, instead, MLD values are a climatology of March MLD monthly means obtained from gridded temperature and salinity data through a density threshold method."

New lines in the discussion (replacing old lines 698-722; now lines 781-785):

"In the LS the wide range of observation-derived MLD estimates, with values of 1000–1200 m in ARMOR3D and 2000–2200 m in EN4, makes model assessment challenging. Analyses from additional observational studies show values between 1100–1500 m (Yashayaev and Loder, 2016) and down to 1400–1800 m (Holte et al., 2017) in the LS, suggesting that HR-HIST mean values for the LS (1800–2000 m) might be slightly too deep, and LR-HIST mean values (1000–1200 m) slightly too shallow."

L355 For this 'distinct stripe', is there any evidence whether it's real or an artifact of the method/data? Is it discussed in other studies?

To our knowledge, it is the first time that a 3D temperature and salinity dataset at the resolution of the ARMOR3D dataset (1/4º) is used to derive MLD estimates. Typically, coarser resolutions of 1º are used, which cannot resolve that feature. The 'distinct stripe' is most probably linked to the resolution of the observational dataset. We have modified the text as follows:

New text (at new lines 383-386):

"A remarkable feature in the ARMOR3D dataset is a distinct stripe of deep mixing (1200–1400 m deep) attached to the shelf along the East Greenland Current, which is also slightly visible in some of the individual HR-HIST models (e.g. HadGEM3-GC31-HH; Fig. 7) and might be absent in EN4 due to its coarser resolution."

Old text:

"A remarkable feature in the ARMOR3D dataset is a distinct stripe of deep mixing (1200–1400 m deep) attached to the shelf along the East Greenland Current, which is also slightly visible in some of the individual HR-HIST models (e.g. HadGEM3-GC31-HH; Fig. 7)."

Fig 12 I know from looking at the BSF in CMIP6 models that the way they are often calculated means that there is a large offset from integrating from the Southern Ocean northwards. You should comment on this somewhere and how you removed this offset since it affect the zero line.

BSF model data in our analyses were directly downloaded from the Earth System Grid Federation portal (ESGF; we did not perform the calculations). Some of the ESGF CMIP6 models in the LR-HIST ensemble presented unrealistic values for the BSF symptomatic of a problem in the integration and/or of differences in the integration approaches. Those models were discarded from the analyses. We have added a figure below (Fig. W) that we produced during the selection of the BSF models, which includes all the models that were downloaded

from the ESGF nodes (also the ones showing problems). That figure is just a snapshot from year 1980 but it shows which were the models that presented problems/differences. There is no evidence of any significant offset in the calculation of the global BSF (including in the Southern Ocean area) in the models kept in our analysis (Fig. W; Fig. 12). Also please note that the CMIP6 models kept in our analysis present a realistic structure for the BSF in the North Atlantic, consistent for example with the SST and SSS biases in the Gulf Stream and CNA area as well as with the AVISO SSH 0 contour line (Fig. 12). A note commenting on these points has been added to the text.

New text (at new lines 148-153):

"Some of the LR-HIST models, as downloaded from the Earth System Grid Federation ESGF data portal, present unrealistic values for the barotropic streamfunction (BSF), reflecting problems in the integration and/or different integration approaches (not shown). Such models have been discarded from our analysis. No significant offset in the BSF (including in the Southern Ocean region) is observed between the models kept in our BSF analysis and they all present a realistic BSF structure in the North Atlantic. Differences originating from different integration assumptions are thus expected to be small in our restricted model sample."

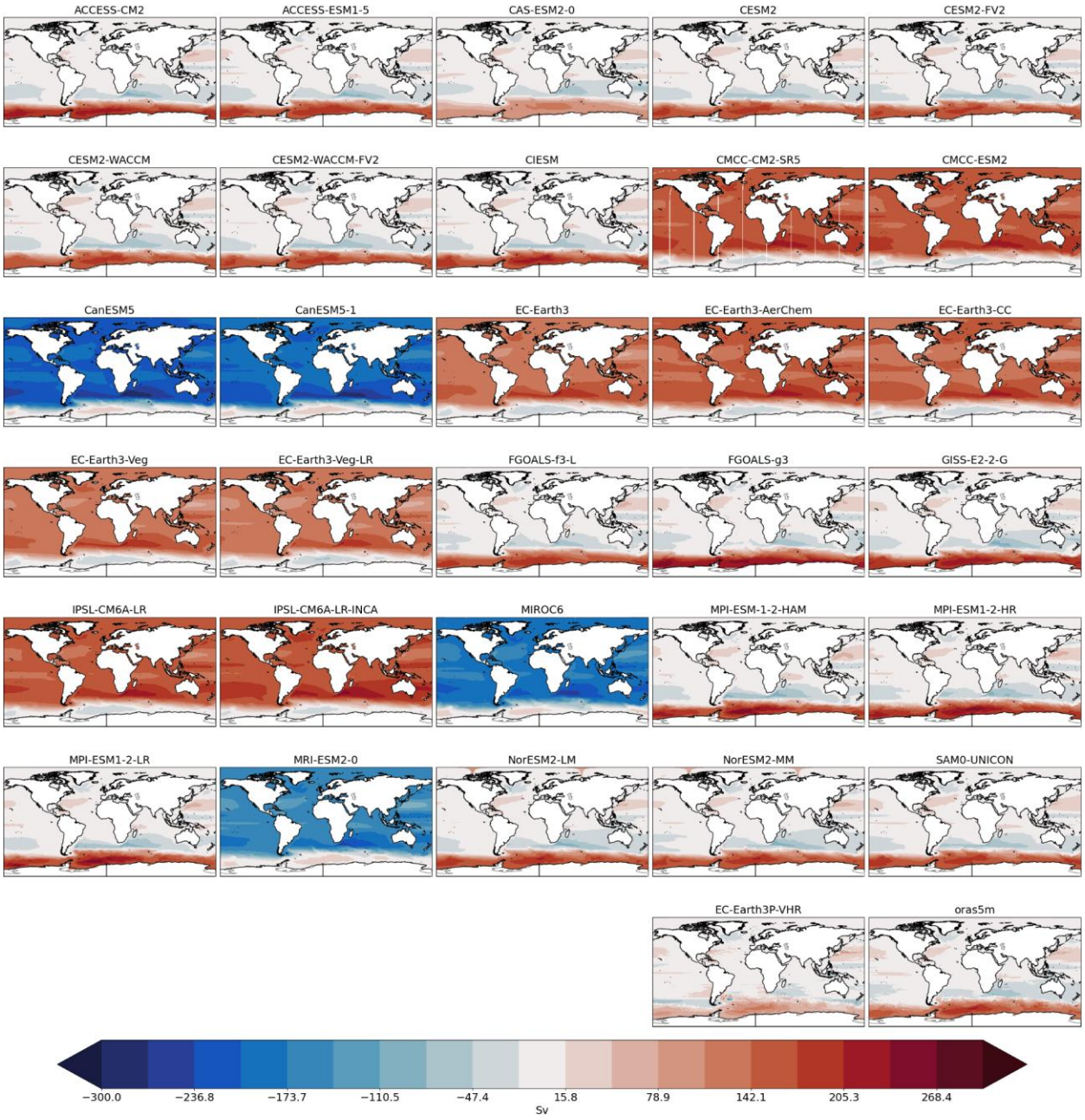

Figure W. BSF in models available from ESGF (year 1980).

L573-576 Fig 14a shows a huge amount of vertical scatter and, by eye, does not look significant to me. Could you please check the correlation and p value for this relationship? Also you quote a correlation from another study. Comparing correlations is pretty meaningless unless you take into account the number of data points. P values are more meaningful to compare.

The exact p-value associated with the correlation value r=0.47 in our calculations is p=0.0019. We observe that models with a weak AMOC tend to have weak convection in the Labrador Sea,

and the other way round. As the reviewer points out, there is some vertical scatter, which is reflected in a moderate correlation value of only r=0.47. The study by Li et al. (2019) does not present the p-values associated with the correlation values. Their correlation values are based on only 5 different data points, so they might be less significant than ours. Please note though that we relate the differences in the r values between the two studies to the smaller size of the model ensemble in Li et al. (2019).

L577 'in consistence' -> consistent

Changed (now line 624).

Section 3.8 It would also be good somewhere to discuss observational agreement – ie that models can have good agreement with multiple metrics at once. E.g. Models with no LS SSS bias can have no AMOC bias etc. A previous study (Danabasoglu, 2014 https://doi.org/10.1016/j.ocemod.2013.10.005) found that although there was a correlation between AMOC and MLD, that models could only agree with one or the other observational constraint.

We have added the following text at the end of old line 581:

New text (at new lines 628-632):

"Note that although the correlations described here are consistent, this consistency does not always translate into full observational fidelity, meaning that some models may agree with observational constraints for one metric but not for the other. For example, whereas several models with max. AMOC values within the range of observational estimates display max. MLD values which are also close to observations, some others exhibit MLD values that are too large compared to observations (Fig. 14a)."

We would like to note though that our study focused on the analysis of the performance of high-resolution models. The observational agreement of each specific high-resolution model has been discussed individually in Section 3.7: Characteristic features in the HR-HIST models as well as in the Discussion section of the manuscript.

L661 'overly weak NAC' – you haven't measured NAC strength only position. Why do you say it is weak?

Figs. 12 and 13 show a weaker barotropic streamfunction (BSF) in the MPI-ESM1-2-ER and EC-Earth3P-VHR models in the CNA box compared to ORAS5m and GLORYS12, as well as a

weaker eastward penetration of the NAC (as represented by the BSF) in these models. The text has now been modified to make these points more explicit.

New Text (at new lines 715-719):

"However, our bootstrapping analysis indicates that the reduction in the CNA surface biases is not statistically significant in our HR-HIST ensemble (Sect 3.6), as some of the HR-HIST models (MPI-ESM1-2-ER and EC-Earth3P-VHR) still present an overly weak NAC in the CNA (as represented in the BSF), with a reduced eastward penetration compared to reanalyses (Figs. 12, 13; Sect 3.7)."

Old Text:

"However, our bootstrapping analysis indicates that the reduction in the CNA surface biases is not statistically significant in our HR-HIST ensemble (Sect 3.6), as some of the HR-HIST models still present an overly weak NAC (MPI-ESM1-2-ER and EC-Earth3P-VHR; Sect 3.7)."

L689-693 Some studies also suggest that the NAC position is important for LS biases – see Jackson et al 2023, Marzocchi et al 2015, Treguier et al 2005 https://doi.org/10.1175/JPO2720.1 This seems relevant here since you've shown the NAC position improves at higher resolution.

Old lines 686-689 have been modified to reflect this.

New text (at new lines 744-748):

"Our study hints that LS and CNA biases might be actually related to each other through northward salinity/heat transport by the NAC, as supported by the correlations between the SSS (and SST) biases of those two regions. Note that northward transports depend both on NAC strength as well as path (Jackson et al., 2023). Studies such as Chang et al. (2020) and Roberts et al. (2019) report increased heat transport by the AMOC in mesoscale eddy-resolving models, further supporting the idea of increased northward transport as a potential origin for the LS biases."

Old text (old lines 686-689):

"Our study hints that LS and CNA biases might be actually related through northward salinity/heat transport by the NAC, as supported by the correlations between the SSS (and SST) biases of those two regions. Studies such as Chang et al. (2020) and Roberts et al. (2019) report

increased heat transport by the AMOC in mesoscale eddy-resolving models, further supporting the idea of increased northward transport as a potential origin for the LS biases."

L761 Should this be 'western continental boundaries'? I don't think it includes those in the east Atlantic.

We have checked Fig. 13, showing BSF multi-model means for HR-HIST and LR-HIST, and we observe that BSF contour lines are closer to each other in the HR-HIST mean compared to the LR-HIST mean, from the western tip of Iceland to the west, which indicates a narrower and locally stronger boundary current. Since these longitudes cover most of the SPG longitudinal extent, we would prefer keeping the original phrasing.

L775-780 This paragraph does not seem relevant to the mean state (the subject of the paper). There might be impacts on the variability/forced evolution, however there are many other studies that would also be relevant then.

This paragraph has now been removed from the text.