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.

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. 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. 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)). 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).

- ***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. 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. 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. 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.

- ***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. 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.

Specific Comments

Abstract

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

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.

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).

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.

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.

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? 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).

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.

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?

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?

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).

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?

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).

Lines 293-294: When comparing to 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?

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?

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.

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.

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).

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.

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.

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.