Referee #1

We would like to thank the reviewer for the constructive comments. Our responses are given below in black type (reviewers comments in blue italics).

Q1.A) It is tacitly assumed that open water deep convection by the parameterization is what is needed to more accurately simulate the Labrador Sea and other deep convection sites. However, there is substantial observational and theoretical evidence that convection influenced by lateral boundaries is more important for the overturning mass flux (e.g., https://doi.org/10.1175/2011JCLI4130.1, https://doi.org/10.1175/JPO-D-24-0019.1).

The academic approaches developed and used by Spall (2010, 2011, 2024) have shown that downwelling is concentrated near the boundary where eddies break away from the current. Therefore, buoyancy-induced downwelling in the Labrador and Irminger Seas, which drives the North Atlantic meridional overturning circulation, is concentrated along their eastern boundaries and not in the center of the basin.
In essence, Spall (2011) showed that coastal downwelling sets up to maintain mass conservation and the balance of thermal winds in the incoming and outgoing flows across the basin.
Based on a vorticity budget in an idealized configuration of buoyancy-driven subtropical and subpolar gyres, Spall (2024) showed that the downwelling near the western boundary results primarily from the balance between the stretching of planetary vorticity and the viscosity induced by eddies, and to a lesser extent from the meridional advection of planetary vorticity. In thermodynamic terms, the net downwelling (W) was expressed as a function of the temperatures of the incoming flow (Tin), the outgoing flow (Tconv), and the gyre center (T0), where deep convection occurs.
Typically, this expression $W \propto (Tin - Tconv) / (Tin - T0)$ shows that the downwelling depends on T0, which is the temperature controlled by the loss of buoyancy at the surface and, ultimately, by convection.
In his simulations, Spall (2010, 2011, 2024) increases the vertical diffusion coefficient (Kz) to 1000 m^2/s for statically unstable conditions to represent vertical convection.
The EDMF parameterization aims precisely to replace this increase in Kz with a mass-flux scheme in order to accurately represent convection and, ultimately, to more realistically simulate the temperature T0 in convection zones, and ultimately the coastal downwelling.

Q1.B) I think you may be missing a major source of convection in under-resolved "Ekman buoyancy flux" that is known to induce strong submesoscale convection (https://doi.org/10.1175/JPO2737.1). This can be included into convective parameterizations straightforwardly (http://dx.doi.org/10.1016/j.ocemod.2016.12.003). This lateral transport can overwhelm the surface buoyancy fluxes in the presence of strong fronts (e.g., https://doi.org/10.1016/j.dsr2.2013.02.025).

Thomas and Lee (2005) and Thomas (2008) showed that potential vorticity (PV) destruction can occur at density fronts under down-front winds. Under such conditions, the Ekman transport advects denser water over lighter water, thereby destabilizing the upper ocean (Thomas and Ferrari, 2008). The associated frictional PV flux, which scales as dTau/dz*DUg/dz, induces an equivalent buoyancy flux (EBF) 10 to 100 stronger than the surface buoyancy loss, as observed at the subpolar front of the Japan Sea (Thomas and Lee, 2005) and along the Ligurian Current in the Western Mediterranean Sea (Giordani et al., 2017). EBF results from the interaction between surface winds and currents, and can

directly destabilize the ocean upper layers and induce vertical mixing by diffusion and convection. The EDMF parameterization does not explicitly diagnose EBF forcing, but responds to its effect on the local density stratification (Giordani et al., 2017). Therefore, when EBF-induced processes destabilize the water column and generate convective conditions, their effects can be represented by EDMF through its response to the resulting unstable stratification.
Following this comment and the previous comment Q1A, we now clarify in the manuscript (lines **162-164)** that EDMF responds to convective instability regardless of its origin, including that induced by surface cooling or salinification, EBF, or coastal downwelling.

Q1.C) Figures 2&3 lack a "control case". Outside of the EDMF parameterization, it is not clear how to interpret the w_p. I suggest comparing the vertical diffusive, total, and convective heat fluxes in the time window shown in Fig. 2 to the vertical total heat fluxes in the EVD simulation at the same location and time. I guess the point is that the MF parameterization will go deeper than the ED flux, but what fraction of the total is it, and is it deeper than in the EVD simulation? In examining KPP variants, we often find that the "deeper" schemes end up being similar to the "shallower" versions under the same surface forcing--it is quite hard to do this comparison fairly (see http://dx.doi.org/10.1175/JPO-D-22-0107.1 and http://dx.doi.org/10.1029/2019MS001810), especially so in a GCM with feedbacks present. Figure 9 does this comparison more effectively than the earlier figures.

We thank the reviewer for this suggestion. As the model outputs required to compute the diffusive and convective heat fluxes at hourly frequency, such as w_p shown in Fig. 2a, are not available, we instead computed the temperature trend from the divergence of the diffusive (∂zF_{diff}), convective (∂zF_{conv}), and total (∂zF_{tot}) vertical heat fluxes from the daily model outputs. The fluxes were calculated at the model grid cell containing the mooring location and over the same time window as shown in Fig. 2a, as :

$$\partial_z F_{diff} = -\frac{\partial}{\partial z}\left(-K_z \frac{\partial T}{\partial z}\right),$$

$$\partial_z F_{conv} = -\frac{\partial}{\partial z}\left(\frac{\partial MF(T_p - T)}{\partial z}\right)$$

In the case of the EVD simulation: $\partial_z F_{tot} = \partial zF_{diff}$; whereas in the case of EDMF simulation: $\partial_z F_{tot} = \partial zF_{diff} + \partial_z F_{conv}$.

In the figure below we compare the contributions of the convective (∂zF_{conv}) and diffusive (∂zF_{diff}) components to the total flux divergence ∂zF_{tot} in the EDMF simulation, and in terms of intensity and vertical extent for the three simulations, under identical surface forcing.

For the EDMF simulation, ∂zF_{diff} is nearly negligible over most of the water column, except for a weak signal close to the surface during the summers. Consequently, the vertical structure of ∂zF_{tot} is almost entirely determined by the convective contribution, ∂zF_{conv}. This confirms that, in the EDMF simulation, the deep response of the total vertical heat flux is primarily associated with the mass-flux component rather than with diffusive mixing. In terms of vertical extent, the EDMF ∂zF_{tot} reaches greater depths than the diffusive signal in the TKE simulations during the strongest convective

winters. This illustrates that the deeper penetration associated with EDMF results from the convective mass-flux contribution, rather than from an enhancement of the diffusive component.

The TKE+EVD and TKE+EVD+MLP simulations exhibit similar temporal variability in ∂zF_{tot} response, with the main convective events occurring during the same winters, but differ in the depth reached by the diffusive contribution: ∂zF_{diff}signal extends to greater depths in the TKE+EVD than in the TKE+EVD+MLP. This is consistent with the results presented in Fig. 9e, which shows deeper MLDs in the TKE+EVD simulation. This is due to a higher EVD triggering frequency in the TKE+EVD case than in the TKE+EVD+MLP case. Thus, the differences in the vertical extent of the diffusive heat-flux contribution are consistent with the differences in the simulated mixed-layer structure.

The interannual variability in the depth of ∂zF_{tot} is also consistent across the three simulations and agrees well with the variability identified in Figs.2a, 9e. In particular, the relatively shallow convective penetration during winter 2012/13 is apparent in all three simulations. The first winter shown here (2011/12), however, exhibits a deeper penetration of ∂zF_{tot} than that inferred from Fig. 2a. This difference is expected because the two figures do not represent exactly the same spatial sampling: Fig. 2a uses hourly model output at the specific mooring location, whereas the present analysis uses daily model output at the model grid cell encompassing the mooring location. Given the strong spatial variability of convection in the Labrador Sea, differences in the diagnosed penetration depth are therefore expected. This result further indicates that the different parameterizations respond consistently to the same atmospheric forcing while producing different vertical extents of the turbulent mixing response.

These results are now added and discussed in new Figs. 2b and S1, and lines **221-227, 323-335**.

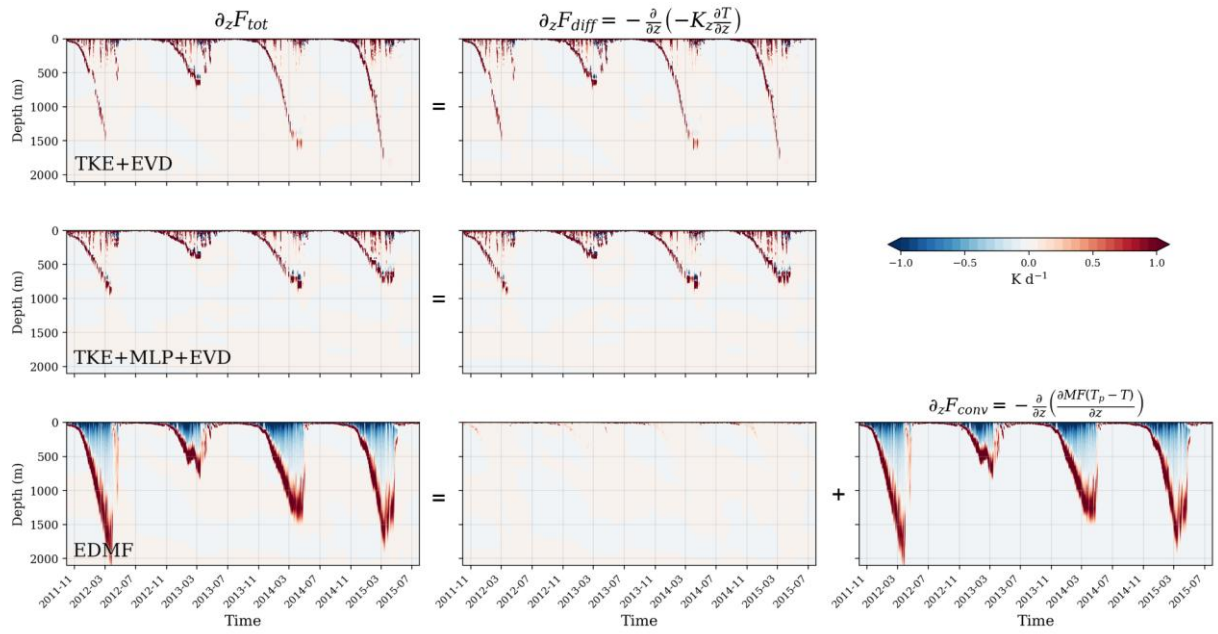


Figure: Vertical divergence (K day^{-1}) of the total (∂zF_{tot}), diffusive (∂zF_{diff}), and convective (∂zF_{conv}) vertical heat fluxes at the model grid cell containing the Labrador Sea mooring (56.53°, 52.65°W), over the same period as shown in Fig. 2a. Rows correspond from top to bottom to TKE+EVD, TKE+EVD+MLP, and EDMF. For the TKE-based simulations, $\partial_z F_{tot} = \partial zF_{diff}$, whereas for EDMF, $\partial_z F_{tot} = \partial zF_{diff} + \partial_z F_{conv}$.

Q1.D) Introduction: A paragraph on the dangers of over-interpreting bias reduction in global models as evidence for improved parameterization physics should be added. The key point is that because of compensating errors in global models, it is often the case that improved physics initially leads to increased bias rather than a reduction. In this case, because of past work on idealized models, care is taken to avoid overinterpretation by comparison to the conclusions drawn in those more controlled settings. Discussion of Figures 5 & 7 should also mention this context, because it is likely that if other parameterizations (e.g., GM) were retuned after implementation of the EDMF scheme, the bias and RMSD reduction might be greater than shown there. Also, you should revise the sentence "Overall, this first global-scale validation demonstrates that EDMF provides a more coherent and physically consistent representation of oceanic convection." correspondingly, as this GCM-bias reduction study does not show coherence nor physically consistent representation, it shows the impacts of those physical principles that are established in the process papers leading up to this one.

We agree with the reviewer that the reduction in biases obtained with EDMF should be interpreted with caution as they reflect the combined effect of EDMF, the other parameterizations, the surface forcing and their interactions. Indeed, the present EDMF simulation was not accompanied by a retuning of other parameterizations following the implementation of EDMF, such as the eddy-induced velocities parameterization (Gent and McWilliams,1990), the submesoscale mixed layer eddies parameterization (Fox-Kemper et al. 2008), or the vertical TKE closure scheme. Further improvements in the global mean state of temperature may therefore be possible through a joint retuning of the model components. We agree that our results should thus be viewed as an indication of the beneficial impact of EDMF formulation, rather than as a standalone demonstration of improved physical realism.

Following these recommendations, we have added a comment regarding the interpretation of bias reduction in global models lines **216-219**, and added/modified lines **751-769, 767-768**.

Q1.E) I'm confused by how large the JJASO bias in the global MLD is for the EDMF scheme vs de Boyer-Montegut, as it seems smaller than that in all of the 3 regions. In the global comparison only, the TKE+EVD seems to be closer to observations during JAS, while in all of the three regional panels the EDMF is closer. Also, you are showing the deB-M and Holte MLDs as though both are comparable to the model, but doesn't NEMO use the deB-M definition of MLD as suggested in Treguier et al. (2023; http://dx.doi.org/10.5194/gmd-16-3849-2023)?

Thank you for pointing this out. Regarding the first part of the comment, we carefully checked the MLD calculation and plotting procedures and found that the MLDs were not area-weighted. We have corrected this issue and produced a new **Fig. 8** using the correct are-weighting procedure for all MLD datasets. As a result, the previously observed large EDMF bias in JJASO relative to de Boyer-Montegut is substantially reduced.
Regarding the second part of the second part of the comment, we confirm that the MLD in NEMO is computed using the density threshold of 0.03 kg m^{-3} relative to 10 m depth, which is the criterion used in the de Boyer Montégut dataset (2004, 2024). We initially included the Holte et al. (2017) datasets to illustrate how the choice of the observational MLD reference dataset can affect the interpretation of the model results.
The “threshold MLD” provided by Holte et al (2017), is computed from individual profiles using the same density threshold proposed by de Boyer Montégut et al. (2004), here as “a density jump equivalent to a temperature difference of 0.2°C at the profile location”. This ultimately leads to a smaller threshold for very cold, dense mixed layers. In addition, this Holte et al. (2017) dataset is

binned into 1°x1° grid cells, without applying a climatological interpolation to fill cells where the number of available profiles is insufficient. Finally, this dataset includes neither the Arctic, the Greenland Sea, the Baltic nor any ice-covered region. In light of these characteristics, and following the recommendations of Treguier et al (2023), we decided to remove the two Holte datasets from the comparison, as their non-global coverage makes them less suitable for comparison with global ocean models than objectively analyzed datasets such as that of de Boyer Montégut et al. (2024).
The text in lines **244-245**, and **533-542** has been revised accordingly.

Minor issues:
L67: Please clarify if these are vertical or horizontal length scales intended, and whether the relevant phenomena are expected to be approximately isotropic or not.

This comment is addressed lines **67**, **69-70**.

L75: It is a bit peculiar to cite the Large et al. 1994 paper for only the diffusive part of the transport, as the majority of the discussion in that paper is about the nondiffusive part of the KPP scheme. The distinctions between the KPP non-diffusive/convective parameterization and the EDMF approach taken here should be delineated nearby.

This is now corrected and addressed lines **77-78**.

L93: author's -> authors' (there is more than one author here)

Correction added line **95**.

L120: are any submesoscale parameterizations used? Are any other parameterizations known to impact ML properties (e.g., penetrating radiation, low cloud physics, etc.) used? These should be laid out.

We thank the reviewer for this comment, more details are provided lines **123-125**.

L140: Please clarify which of these papers is the relevant first source for the MLP parameterization by adding the reference to the sentence ending on L133.

Correction added line **137.**

L166: Why are you using rho for pressure and for density? That is really confusing. Please use p_h for hydrostatic pressure if that symbol is correctly defined in the text.

Correction added to Eq. 1 and in corresponding text.

Eq (2) and subsequent. I am very confused by the dimensions of the properties in this equation and the related notation. Cursive L (suggesting a length) is an inverse length scale while epsilon and delta are described as entrainment "rates" when indeed they are also inverse length scales, as the beta parameters are dimensionless. Please clarify that all of these are the vertical decay scale for the exponential functions, which may increase area (if entrainment plus slowing dominates) or decrease area (in the case of dominant detrainment). The slowing rate (cursive L) is meant to capture the

cross-sectional momentum adjustment of the plume as it widens in the absence of entrainment or detrainment.

This is now addressed lines **183-185**.

L200: this comment about T_p vs. T at the second half time step makes it quite confusing what the implied flux of heat, salt and freshwater are in this scheme. Please clarify how these are set at the surface for both the diffusive and mass flux portions of the scheme so as to ensure conservation.

This is now clarified lines **211-212**.

Figure 10 caption should mention that the sign convention for the Delta metrics is given in Equation (12). Also, Panels a&b should have 2 y-axis labels, one in temperature units for Delta T and one in W/m^2 for SURF. I have no idea what W/m^2 numbers are given the normalization of the tables and the lack of this second vertical axis label. Ideally, SURF in c & d should be given in both temperature and W/m^2 units.

Following the reviewer suggestion's, Fig. 10 and caption are modified accordingly.

Referee #2

We would like to thank the reviewer for the constructive comments. Our responses are given below in black type (reviewers comments in blue italics).

The manuscript presents a global ocean modelling study using the NEMO framework to evaluate the performance of an Eddy-Diffusivity Mass-Flux (EDMF) parameterization for vertical mixing. The study compares EDMF with traditional eddy-diffusivity-based schemes (e.g., EVD and TKE-based closures) in simulating ocean temperature, salinity, and mixed-layer depth (MLD), with a particular focus on convective processes across different regimes.
The work is timely and relevant. Representing convection and nonlocal mixing remains a major challenge in ocean and climate models, and the implementation of a mass-flux framework inspired by atmospheric convection is a promising direction. The manuscript is generally well structured and clearly presents the modelling framework, experimental setup, and evaluation strategy. I have identified several major and minor issues that the authors should address. Addressing these points will significantly improve the clarity, physical interpretation, and overall quality of the manuscript.

Q2.A)
General Comment (Convective Regimes): The manuscript does not clearly distinguish between different convective regimes (e.g., shallow tropical vs deep polar convection), despite their fundamentally different plume dynamics and forcing. It remains unclear whether the EDMF formulation appropriately captures these regime-dependent processes or applies a uniform parameterization across all conditions. Please clarify how the scheme represents these contrasting regimes and whether its performance varies between them.

In the Introduction (lines **49-59**), we provide definitions of the two convective regimes in terms of their dynamics and forcing. In section ***2.2 Vertical parameterization options***, we present the equations

and coefficients of EDMF, which are applied uniformly across convective regime. The scheme does not use regime-specific formulations; the resulting convective fluxes vary with the local stratification and surface forcing, leading to different convective responses under shallow-diurnal and deep-seasonal conditions.
We then assess the model performance separately for the two contrasting convective regimes in sections ***3.1.1 Deep-seasonal convection*** and ***3.1.2 Shallow-diurnal convection***. Using model outputs at 1-hour resolution, we show that EDMF simulation captures the timing, penetration depths and intensities of both convective regimes. The ability of EDMF to reproduce these two distinct regimes is further summarized in section ***5 Discussion and Conclusion***, lines **730-738**. A quantitative assessment of whether EDMF performs better in one convective regime than in the other would require long-term, high-resolution observational datasets identical for both regimes, which are currently not available.
We have now clarified in the manuscript that the EDMF parameterization uses a unique formulation across convective regimes, while its response to local forcing and stratification allows it to represent both shallow diurnal and deep seasonal convection, lines **160-163**, **726**.

Q2.B) Finally, the manuscript does not discuss the role of double-diffusive convection, which can be important in stratified regions of the ocean where temperature and salinity gradients compete. It is unclear whether the EDMF formulation is capable of representing double-diffusive processes (e.g., salt fingering or diffusive convection), or whether these effects are entirely neglected. Please clarify whether the scheme accounts for double-diffusive mixing and include a brief discussion of its potential impact on the results.

In the ocean, double-diffusive convection is a phenomenon in which water movements are generated by the difference in diffusion rates between heat and salt. Heat diffuses (Kh) approximately 100 times faster than salt (Ks). This disparity creates “salt-fingers” or “thermohaline steps” types of instability, which occur within stable regimes. EDMF inherently accounts for the double-diffusion phenomenon through its ED component, where distinct values for Kh and Ks can be specified. However, under unstable conditions, deep convection simultaneously transports heat and salt during turbulent mixing. As a consequence, there is no separate convective vertical velocity (w_p) for heat and salt. We have now addressed the issue of the interaction between convection and double-diffusion, lines **777-780**.

Specific comments :
Q2.C) Fig. 4: Figure 4 shows a distinct structure around ~60 m depth that is not captured by any of the models. This feature may be influenced by processes such as inertial oscillations or internal waves, which are not explicitly represented in the current parameterizations. Please clarify the physical origin of this structure and discuss whether its absence reflects limitations of the vertical mixing schemes or missing wave-driven processes.

We suspect that these structures are associated with tidal variability, which can generate oscillations near the thermocline. Explicit tides are not represented in our simulations, and this missing process may therefore contribute to the inability of the model to reproduce the observed structures. This point has now been clarified in the revised manuscript (lines **400–401**).

Q2.D) Fig. 5: Figure 5 reveals substantial and systematic temperature biases in the subsurface (thermocline/pycnocline) across all parameterizations, suggesting overly strong vertical mixing below the mixed layer and an unrealistic downward transport of heat that weakens stratification. Given the central role of the thermocline in regulating heat storage and vertical exchange, it remains unclear whether the EDMF scheme provides a physically meaningful improvement in vertical

structure, despite improvements in mixed layer depth. Please clarify the origin of these subsurface biases and discuss whether they arise from excessive mixing across the base of the mixed layer or from missing ***restratification processe****s, as addressing this would strengthen the overall interpretation of the results.*

The comparison between parameterizations suggests that the subsurface (~ 50 m depth) cold biases cannot be attributed systematically to excessive mixing below the mixed layer (Fig. 5c,e,g). In some cases, increasing turbulent transport below the mixed layer actually reduces the bias.

In the tropical ocean, both EVD simulations exhibit similar cold subsurface bias structures, although the addition of the MLP parameterization reduces the magnitude of the maximum bias by approximately 0.25 °C (Fig. 5c). Since the MLP parameterization introduces an additional TKE source below the mixed layer, the reduction of the cold subsurface bias in the TKE+MLP+EVD simulation suggests that insufficient vertical mixing below the mixed layer contributes to the tropical subsurface bias, rather than excessive mixing. By increasing the penetration of turbulent mixing below the mixed layer, MLP allows a greater downward redistribution of heat and consequently reduces the cold subsurface bias.
A similar mechanism can explain the improvement obtained with EDMF. By allowing the penetration of mixed warm waters to greater depths during nighttime, and therefore enhancing nighttime destratification to greater depth (Fig. 4), EDMF increases the downward extent of convective mixing and reduces the subsurface cold bias (Fig. 5c). The daytime restratification is comparable in the three simulations, despite their different nighttime destratification depths (Fig. 4b–d). Therefore, the comparison does not provide evidence that differences in daytime restratification between the simulations are the primary cause of the tropical subsurface cold biases.

A similar behaviour is found in the northern and southern hemisphere oceans (Fig. 5e,g), although the magnitude and vertical structure of the biases differ from those in the tropical ocean. In both regions, TKE+EVD exhibits a cold subsurface bias, which is also present in the EDMF simulation. The addition of the MLP parameterization substantially reduces this cold bias, indicating that the additional TKE source below the mixed layer can improve the representation of vertical heat redistribution in these regions as well. This result is again more consistent with insufficient, rather than excessive, turbulent mixing below the mixed layer as one contributor to the subsurface cold bias.
However, the comparison (Fig. 5e,g) also shows that increasing mixing below the mixed layer does not systematically lead to a better representation of the subsurface temperature structure. In the northern hemisphere ocean, the additional mixing associated with MLP largely compensates the cold bias and produces a warm bias (Fig. 5e), suggesting that the enhanced turbulent transport may become excessive at some depths. This highlights that insufficient mixing can contribute to the cold bias, whereas excessive mixing can result in an over-compensation.
For the northern and southern Hemisphere oceans, we do not have diagnostics that allow us to quantify the contribution of restratification to the remaining biases. Therefore, we cannot exclude a contribution from differences in the balance between mixing and restratification in these regions.

Overall, these results suggest that the subsurface temperature biases cannot be attributed systematically to excessive vertical mixing below the mixed layer. In the tropical ocean in particular, the reduction of the bias when additional turbulent mixing is introduced below the mixed layer indicates that insufficient mixing contributes to the cold subsurface bias.

In any case, the goal of EDMF is to move away from EVD and MLP tricks in order to physically represent turbulence at the base of the mixed layer. We recognize that this is a challenging task, as the model was originally tuned using EVD and MLP. Furthermore, we know that convection interacts strongly with eddy-induced restratification. In the future, we plan to use the GEOMETRIC parameterization to represent sub-mesoscale restratification in the NEMO code. Consequently, the next step will naturally involve parameter tuning using EDMF, GEOMETRIC, and likely other new parameterizations. Text was added lines **764-771**.

Q2.E)Fig. 7: The manuscript does not adequately discuss model performance in western boundary current regions (e.g., Gulf Stream, Kuroshio), where noticeable MLD biases appear in Fig. 7. These regions are strongly influenced by lateral processes (eddies, fronts, advection), and not solely by vertical mixing. Please clarify whether the reported differences with EDMF in these regions reflect improvements in vertical physics or are influenced by unresolved horizontal dynamics at the given resolution. A brief discussion addressing this point would strengthen the manuscript.

We thank the reviewer for highlighting the specific challenges associated with western boundary current (WBC) regions. We already mention in the manuscript (lines **485–487**) that both vertical mixing parameterizations exhibit similar warm temperature biases in these regions. These biases are likely related, at least in part, to deficiencies in the representation of low-level clouds in state-of-the-art climate models, which affect the associated longwave and shortwave radiative fluxes (Zhang et al., 2023).
At the global scale, the difference in temperature biases between EDMF and the two EVD-based simulations is predominantly negative when considered in absolute terms (Fig7i,j,k), indicating an overall reduction of temperature biases with EDMF. However, positive differences remain locally in the Gulf Stream and Kuroshio regions. These residual biases should be interpreted with caution, as WBC are strongly affected by horizontal processes, including mesoscale eddies, fronts, and lateral advection, which are not fully resolved at the 1/4° horizontal resolution of our simulations. Consequently, the remaining positive anomalies do not necessarily indicate a deficiency of the vertical mixing scheme, as they may partly reflect errors in the representation of horizontal dynamics and air–sea interactions.
Further calibration of the EDMF parameters may improve the overall model performance in WBC; however, tuning the scheme specifically to compensate for biases associated with the unresolved dynamics of WBC could lead to over-compensation and a loss of physical consistency elsewhere. We have therefore added a discussion of this limitation in the revised manuscript, lines **487-491**. This point is also consistent with our response to Reviewer #1, Q1.D, regarding the need to avoid over-interpreting reductions in model biases as evidence of improved physical realism.

Q2.F) Justification of MLD criterion: The authors should (i) justify the use of the Holte et al. method more explicitly, and (ii) assess the robustness of their conclusions by comparing with at least one alternative MLD definition (e.g., a fixed density threshold). This would strengthen confidence that the reported improvements with EDMF are physically meaningful rather than diagnostic-dependent.

Following the reviewer's comment and that of Reviewer #1 (see our answer to Q1E), we decided to remove the comparison to the Holte et al. (2017) datasets, which are not adequate for comparisons with climatology of global circulation models. Please see our new Fig. 8 and text lines **526-536**.

Q2.G) Fig. 8: While Fig. 8 demonstrates improved seasonal MLD variability, this metric alone does not uniquely validate the underlying plume dynamics. Similar improvements could arise from compensating errors in parameter tuning. Additional diagnostics (e.g., vertical buoyancy flux, entrainment rates, or plume energetics) are required to confirm that EDMF captures the correct physical processes.

In the present study, the three simulations use the same model parameters, with the exception of those defining the respective vertical mixing parameterizations. Thus, the differences between the simulations arise from their different representations of vertical mixing rather than from differences in parameter tuning. Following this comment and that of Reviewer #1 (Q1.C), we have added Fig. 2b and Fig. S1, which provide additional diagnostics of the vertical heat fluxes. Figure 2b separates the diffusive and convective contributions to the total vertical heat flux in the EDMF simulation, thereby directly illustrating the contribution of the mass-flux component to vertical heat transport. Figure S1 compares the intensity and vertical extent of the total vertical heat fluxes across the three simulations under the same surface forcing. These diagnostics provide additional evidence that the improved MLD variability in EDMF is associated with a distinct representation of convective transport, rather than solely with compensating errors from parameter tuning. Please refer also to lines **214–220** and **316–339**.

Q2.H)Fig. 9: In Fig. 9f,g, only model results are shown without comparison to observational data. Given that these panels appear central to interpreting the behavior of the EDMF scheme, it is unclear how well the simulated patterns reflect reality. Please clarify why observational data are not included for these diagnostics, and whether suitable datasets exist for comparison. Including or discussing observational constraints would strengthen the credibility of the results.

We thank the reviewer for this comment. We agree that observational datasets are important for assessing the realism of the simulated MLD. However, the purpose of Fig. 9 is not primarily to demonstrate that EDMF produces a more realistic MLD than the other parameterizations, but rather to provide insight into the mechanisms underlying the differences in MLD.
Observational estimates of MLD are available in the Labrador Sea. For example, Holte et al. (2017) provide an Argo-based MLD climatology as well as MLD estimates for individual Argo profiles. However, these observations are not directly comparable to the quantity shown in Fig. 9. Our simulated MLD time series are calculated as spatial averages over a box encompassing the mooring location and extending over both the deep Labrador Sea and shallower regions over the continental slope. In contrast, Argo observations provide discrete, irregularly distributed measurements at individual locations and times, with the observational sampling being strongly heterogeneous in space and time. In particular, the Argo array samples the deep basin (depths > 2000m), while the shallower regions of the continental slope are not sampled. Consequently, directly comparing the model box-averaged MLD with an Argo-based time series would require a dedicated space- and time-collocation of individual Argo profiles with the model output, including the computation of MLD from each collocated model profile using the same criterion as applied to the observations. We considered that this detailed model–observation comparison was beyond the scope of the present study.
Nevertheless, the Argo-based results of Holte et al. (2017) provide an independent observational constraint on the seasonal and interannual variability of MLD in the Labrador Sea, and we now discuss this observational context in the revised manuscript. We have also clarified in the text (lines **579-583**) that Fig. 9 is intended primarily as a diagnostic of the mechanisms responsible for the differences between the parameterizations, rather than as a direct assessment of the realism of the simulated MLDs.