DETAILED RESPONSES TO REVIEWER #1

For completeness, we suggest reviewer #1 to read our answers to reviewer #2, since some of the remarks are overlapping.

Overall: The present study investigates intrinsic variability of the Eighteen Degree Water (EDW), the subtropical mode water of the North Atlantic using a 50-member ensemble simulation with 1/4-degree horizontal resolution, eddy-permitting ocean general circulation model. The new method of estimating the surface heat flux for each ensemble member is introduced to avoid artificially damping the intrinsic variability. This new method is interesting. However, I have several questions and comments, as mentioned below, that should be addressed.

Recommendation: major revision

Main comments/questions:

1. The method for estimating surface heat flux of each member and its influence

As shown in Figure A1, this method of surface heat flux strongly affects the amplitude of intrinsic variability, which is the main topic of the present paper. Then I think that meaning of the method and its influence should be discussed further.

1-1. Although it is not clearly mentioned, I guess that this ensemble simulation with the same time-varying air-sea fluxes is original of the present study. If it is not the original of the present study, please add the reference(s) in the description of the method.

This is indeed the first time this specific ensemble simulation is described and used in a publication. We apologize for a confusing statement on this point in footnote #1 in page 3 (numbering of the previous manuscript) : rather than what was written, the present ensemble simulation (NATL025-GSL310) is close but not identical to the E-NATL025 simulation described in Bessières et al (2017). The corrected footnote now reads :

"*This regional ensemble simulation, referred to as NATL025-GSL301 in the OCCIPUT database, is similar to the E-NATL025 simulation described in Bessières et al. (2017) with two differences: its size (50 members instead of 10) and its atmospheric forcing function, as described below. The technical implementation of OCCIPUT ensembles is described in detail in the latter paper.*"

1-2. I think Appendix A should be included in the main text.

We agree: now section 2.1 is split into 2 subsections.

1-3. "as would be expected (i.e. with no excessive damping) in coupled ocean-atmosphere simulations" (line 109). I agree that it would be expected if the atmosphere has enough time to respond to SST sufficiently. However, it should be noted that observational data show enhanced upward surface heat flux over warm meso-scale eddies (e.g., Tomita et al. 2019, doi:10.1007/s10872-018-0493-x), implying that oceanic intrinsic variability might modify surface heat flux, as meso-scale eddies could be expected as intrinsic variability. Then I think that the actual situation is between the two methods of ensemble simulations, and this new method might overestimate strength of intrinsic variability as it is not damped by surface heat flux. As this discussion can directly relate to the main topic of the present study, the meaning of the method and this possibility of overestimation should be discussed carefully.

Thanks for this interesting comment. A way to summarize the reviewer's and our main viewpoints on this subject is the following. [1] We agree that the intrinsic variability interacting with air-sea fluxes in the real ocean-atmosphere system may behave in a way somewhere between those in the ensembles driven by "member specific" (MS) and "ensemble averaged" (EA) forcings. [2] However, the reviewer thinks EA forcing "might overestimate the strength of intrinsic variability"

while we think it is not likely, and we think the EA method is more physically-consistent. Let us elaborate on this in more detail.

We fully agree that SST anomalies have a direct impact on turbulent air-sea heat fluxes, both in the real ocean and in this forced ocean simulation. This is true as well for intrinsic anomalies, both at mesoscale (as mentioned by the reviewer) and at larger space/time scales.

In ocean-only simulations with MS forcing, this effect introduces a strong negative feedback that damps SST anomalies (and the SST spread in the MS ensemble) that emerge spontaneously, over timescales of about 30-45 days in this region (Barnier et al 1995). A key point is that such forced ocean-only models behave such as if the atmosphere were totally insensitive to the ocean, i.e. such as if the atmosphere had infinite heat capacity. This is non-physical, since heat capacity is instead much smaller in the atmosphere than in the ocean: this damping timescale is therefore very likely too short in such models, in particular for intrinsic variability (since the "forced" variability has to follow the atmospheric pacing). This is why we argue that the MS forcing technique overdamps and underestimates the SST spread (in particular at interannual-to-decadal timescales on which we focus here), and that intrinsic variability behaves in a more physically-consistent way with the EA forcing approach.

In contrast, it is not obvious to us why the ensemble with EA forcing "might overestimate the strength of intrinsic variability"; this hypothesis is in fact difficult to verify since we have no measurement of it in nature. Instead, the variability simulated by NEMO is often underestimated at 1/4° with the MS (i.e. classical bulk-based) forcing at all scales that have been examined (to our knowledge). This is the case in particular for the mesoscale intrinsic variability that is weaker than observed in such simulations (see Penduff et al, 2010, DOI: https://doi.org/10.5194/os-6-269-2010); this holds as well as for low-frequency intrinsic variability, which is smaller at 1/4° than at 1/12° resolution (Sérazin et al, 2015, [DOI: 10.1175/JCLI-D-14-00554.1], Sérazin et al 2018 [DOI: 10.1175/JPO-D-17-0136.1]). Intrinsic variability is thus underestimated at 1/4° with the MS approach. We argue that its upper ocean enhancement obtained with the EA method goes in the right direction.

We thus agree with the reviewer on point [1] above, but less so on point [2]. To summarize this interesting discussion, we propose to clarify in the new section 2.1.2 the description of the method, and to discuss these considerations, as follows:

"*In other words, using ensemble-averaged instead of member-specific air-sea fluxes does not adversely affect the atmospherically-forced oceanic state and evolution, and enhances the ensemble dispersion of yearly temperatures in the upper 300 m. We explain this latter enhancement and argue that this ensemble-averaged forcing method is preferable, as follows.*

The classical (member-specific) computation of turbulent air-sea fluxes through bulk formulae in ocean-only simulations induces an implicit relaxation of sea-surface temperature (SST) toward a prescribed and fluctuating equivalent air temperature T_a, with a time scale on the order of 40 days in our region of interest (see Fig. 6 in \cite{Barnier1995}). This relaxation is arguably overestimated in such simulations where the heat capacity of the atmosphere is assumed infinite despite its being much smaller than that of the ocean in nature. In an ensemble simulation driven with member-specific fluxes, this results in SSTs being over-relaxed toward the same T_a within all members; this in turns yields an excessive damping of ensemble SST dispersion at these long timescales in particular, and of intrinsic variability in general. Indeed, previous 1/4°-resolution NEMO simulations driven by classical (member-specific) forcing have been shown to underestimate surface intrinsic variability at all scales, compared to observations and to 1/12° simulations \citep[see e.g.][]{penduff2010, Serazin2015, Serazin2018}.

The ensemble-averaged forcing method avoids this excessive damping of surface CIV and lets intrinsic temperature anomalies reach up to the surface. Such a behavior is probably closer to that in coupled ocean-atmosphere simulations, where the ocean's thermal inertia overwhelms that of the atmosphere; estimating the strength of interannual CIV in eddying coupled models would help verify this hypothesis. Nevertheless, the use of ensemble-averaged instead of member-specific

fluxes removes this unphysical imbalance between the oceanic and atmospheric heat capacities, and compensates the lack of simulated intrinsic variability. We hypothesize that the amplitude of upper-ocean temperature interannual CIV in nature sits between those simulated with both forcing strategies, and argue that the ensemble-averaged forcing method lets it evolve in a more physically-consistent and realistic way."

2. Reliability of the observational data product

We thank the reviewer for these remarks. Before answering them in detail below, we would like to make a few comments.

The main goal of the paper is to provide first estimates of the relative contributions of forced and intrinsic interannual variabilities on EDW properties. These relative contributions happen to be mostly consistent with ARMOR3D (see section 3.1.3 in the manuscript) and may be informative for the community, even if the total (forced+intrinsic) simulated variability somewhat differs from the real one; future studies on the subject (potentially using higher resolution ensembles) will help refine our first estimates.

As explained below, the realism of the total simulated variability is evaluated using the only available observation-only gridded dataset (ARMOR3D). Like all observation-based products, this product is not perfect: it tends to underestimate total variabilities by design, and we take this fact into account when interpreting the results.

2-1. If ARMOR3D is not reliable, as discussed around line 218, I think it would be better for the authors to use other observational data product(s) to evaluate the model result.

We do not consider nor mention that ARMOR3D is "unreliable" (nor "reliable"). Let us first list the reasons why we chose ARMOR3D (see section 2.2 of the manuscript).

[1] At relatively coarse scales (e.g. 2-3°), ARMOR3D's multivariate fields are fully constrained by in-situ observations like other OI-based products (like ISAS for instance); in addition, ARMOR3D fields are fully constrained by altimetry, i.e. at eddy scales as well. ARMOR3D thus provides a 3D gridded multivariate U,V,T,S dataset allowing us to compute the full Ertel PV (including relative vorticity) as in the model simulation, at a comparable resolution.

[2] ARMOR3D does not rely on a numerical model; our next point [3] and our second paragraph in 2-2 below explain why we prefer such a gridded product instead of a reanalysis, which is based on a numerical model.

[3] Eddy scales are present in 1/4° or 1/12° reanalyses as well since they are produced by CGMs, but this does not mean they are better constrained than in ARMOR3D.

[4] ARMOR3D is the only existing product of this kind.

For all these reasons, we consider there is no alternative observation-only 3D gridded product that may be substituted to ARMOR3D for full Ertel PV computation.

2-2. If ARMOR3D is not reliable for its amplitude of variability, I seriously wonder if the phase of variability in ARMOR3D is reliable. Please add some discussion on it.

The amplitude and phase are distinct features of the variability in ARMOR3D. Pauthenet et al (Ocean Science 2022) show in their figure 12-a that the interannual large-scale SST fluctuations in ARMOR3D over our region are perfectly in phase with satellite observations and with the GLORYS12 reanalysis; these authors also note that the amplitudes of these fluctuations are very similar in ARMOR3D and GLORYS12. Balmaseda et al (Journal of Operational Oceanography 2015) assessed several reanalyses and observation-based analyses against observed data locally, and did not identify any phase shift of SSH variability in ARMOR3D. In particular, their figure 3-a indicates that ARMOR3D yields the smallest RMSE and the best correlation of all available

products (including reanalyses) with the local SSH (mostly interannual) variability estimated at tide gauges. No hint of phase shift has thus been reported in this dataset, both at large and local scales.

More generally, it would be questionable to use a reanalysis instead of ARMOR3D as an "observation-based" reference with smaller uncertainties: reanalysed products do not only differ from each other, they also depend in complex ways on specific model parameters, resolution, assimilation techniques, forcing fields, etc (e.g. Balmaseda, 2015). Differences between our results and a given reanalysis (GLORYS12, ECCO1°, etc) would depend on these complex dependencies, so that interpreting them would raise more questions than provide answers, and likely bring more confusion than clarity in the assessment of our simulation. The choice of ARMOR3D as a reference avoids these complex issues, this product was shown to be as reliable as several reanalyses (including GLORYS12) in various studies, and it comes with a known information regarding its limitations. ARMOR3D thus appears adequate for model assessment in this region, despite its imperfections.

To summarize the discussion in parts 2, 2-1 and 2-2 above, and to address the reviewer's requests, the revised manuscipt no longer refers to ARMOR3D as "observed" data or "observations" (which was somewhat misleading), but as "observation-based data" (or simply as "ARMOR3D"). We have also moderated several of our statements about the "good" (replaced by e.g. "relatively good" or "correct") model-ARMOR3D agreement at several places in the paper. Finally, we have rewritten section 2.2 of the manuscript in order to clarify our choice, and included a short discussion about the above with a direct reference to Balmaseda et al (2015):

"We use ARMOR3D over its first 34 vertical levels (i.e. down to about 800 m) to assess the model simulation over our region of interest and the whole simulation period. ARMOR3D is a global analysis based on observational datasets including satellite sea surface temperature (SST), altimeter-derived sea surface height, in-situ temperature/salinity profiles from the Argo array, CTD and XBT profiles. These observations were processed to provide temperature (T), salinity (S), and geostrophic velocity (u,v) fields on a 3-D grid at 1/4° resolution using optimal interpolation and multiple linear regression methods as explained in \cite{Guinehut2012} and \cite{Mulet2012}. This latter study presents how gridded T and S fields are used to provide consistent 3-D velocity fields via the thermal wind relation, with a surface reference level where geostrophic velocities are derived from altimetry.

ARMOR3D has some uncertainties and limitations, as any gridded product constrained by observations. Episodic spurious density inversions have been detected in ARMOR3D near the surface (E. Pauthenet, personal communication), but these artifacts do not affect the subsurface where most of the STMW is found. The interannual variability (in particular of salinity) is also known to be somewhat underestimated in ARMOR3D \citep{Guinehut2012}, partly since the coverage of in-situ data is relatively coarse and since optimal interpolation has a tendency to smooth solutions.

The ARMOR3D dataset also has strengths despite its limitations, and it was chosen as our observation-based reference for three main reasons, the first two of which are documented in \cite{Balmaseda2015}: [i] ARMOR3D compares well with independent observations at local and large-scale in our region of interest, with a skill that is similar to ocean reanalyses. [ii] The ARMOR3D fields are independent of multiple and complex modelling choices, which produce substantial differences between reanalyses. [iii] Perhaps more decisively, ARMOR3D is the only available model-independent T,S,u,v dataset that yields the full Ertel PV (including ζ) at a spatiotemporal resolution that is close to that of our model. As in all comparisons between simulations and any observation-based gridded dataset, the specificities of ARMOR3D will be taken into account in the comparisons discussed below."

3. Influence on the annual cycle

Line 195: "the large control exerted by the atmospheric annual cycle" As mentioned here, the annual cycle of forcing is exceptional. It would be interesting to investigate the time-scale dependence of CIV strength excluding the annual cycle.

We thank the reviewer for this suggestion: analyzing the forced and intrinsic fluctuations of EDW at subannual scales would be interesting, and would indeed require a removal of the mean seasonal cycle. This may be the subject of a subsequent, dedicated study. However, the present study is focused on EDW interannual fluctuations, and we accordingly removed the mean seasonal cycles before our analyses (as mentioned in line 160).

Specific comments/questions:

Line 15: "a notable role in regional and global climate" It would be better to explain more explicitly.

We agree. The sentence starting with "It is a weakly stratified…" has been replaced with the following:

"It is a weakly stratified, homogeneous water mass sitting on top of the permanent pycnocline with constant temperature near 18^oC \citep{WORTHINGTON1958297, Feucher2016}. The EDW plays a notable role in climate and ecosystems, most notably because it is a significant heat and anthropogenic carbon reservoir \citep[e.g.][]{Dong2004, Bates2002, Bates2007, Kelly2010, Perez2013} that further supply or deplete oxygen and nutrients to the subtropical gyre and the Western boundary current system \citep[e.g.][]{Jenkins2003, Palter2005} "

Line 52: "EDW" should be STMW, I think.

Thanks for this remark which made us realize that we used the acronym STMW in the title and EDW in the paper: this led us to the question "which of the two names should we use?". The fact that most papers that we cited use STMW instead of EDW made us choose STMW, in the title and everywhere in the text.

Line 117: It would be good if the authors can add a brief comment on how the gridded T and S fields are dynamically consistent with the velocity field.

Thanks for this suggestion. We have added a reference to Mulet et al (2012), a study that complements Guinehut et al (2012) and which describes how dynamical consistency is ensured between T,S and U,V. We have modified the following sentence in the paper as follows:

"These observations were processed to provide temperature (T), salinity (S), and geostrophic velocity (u,v) fields on a 3-D grid at 1/4° resolution using optimal interpolation and multiple linear regression methods, as explained in Guinehut et al. (2012) and Mulet et al (2012). This latter study presents how gridded T and S fields are used to provide consistent 3-D velocity fields via the thermal wind relation, with a surface reference level where geostrophic velocities are derived from altimetry. The multivariate ARMOR3D dataset was chosen as an observation…"

Line 147: Although I know that model simulations always have biases, it is usual to adjust the parameters to define the simulated EDW for comparison with the observed EDW. The authors may want to add some more explanation why they tried to adjust the parameters for the observed EDW.

The ARMOR3D dataset is indeed based on observations, which have been substantially processed: the underlying OI algorithm yields a product that differs from the original data, and thus introduces an uncertainty as to the criteria to identify the mode water (note that there is also some arbitrariness when choosing EDW criteria from a set of CTD profiles). Testing 3 sets of slightly different criteria is a means to evaluate the influence of these uncertainties on ARMOR3D EDW properties. To clarify this point, we have modified the second paragraph of section 2.4 as follows :

"The PV maximum and geographical boundaries select weakly stratified waters in the region of interest, and the density range excludes those located outside the layer located between the seasonal and main thermoclines. The PV maximum and density range and have different values in the model ensemble and the observational product to account for the differences between the observed and simulated ocean states (see Section \ref{subsec:seasStruct}). The ARMOR3D gridding algorithm also yields some uncertainty as to which exact criteria should be chosen to

identify EDW. This uncertainty was evaluated using various sets of values for PV and density: three of these are presented here, defined in Table \ref{tab:EDWhere} as A, B and C, with increasingly larger bounds. Section \ref{sec:Robustness} evaluates the effect of the different values used in setting the boundaries of EDW in both datasets."

Line 162: "latter two sets of" It would be better to describe more specifically.

We agree and we have clarified this sentence.

Figure 4: Please improve the labels of Figure 4. Some of the labels on the panels in the right column are overlapped and cannot be read well.

We agree. Taking into account both reviewers' comments, Figure 4 has been modified to improve readability; the caption has been adjusted accordingly.

Figure 5: As only correlations are discussed in this paragraph, it might be more appropriate to plot only correlations rather than using the Taylor diagram. As sometimes STDs are very different, it is difficult to compare the distribution of correlations in Figure 5.

Thanks for this suggestion. The Taylor diagram has been replaced with pdfs of correlations in the new Figure 5. The caption and text have been adjusted accordingly.

Line 255-256: Although I agree that the range is overlapped, the distributions seem very different, and the discussion here seems not objective. The authors may want to add some more objective and quantitative discussion.

The superimposed distributions of correlations in the new Fig 5 are indeed easier to compare than the previous Taylor diagram. This new figure confirms the existence of overlaps between most distributions (which are particularly clear for thermodynamical variables), except for the depth of EDW (as was stated in the first manuscript). We have slightly adjusted the text based on this new figure and on the reviewer's remark.

Line 291-296: Although it is good to mention here, I do not think the discussion in this paragraph is a new finding of the present study.

We agree that the use of Taylor diagrams and related statistics for ensemble assessment is not new, but [i] very few oceanographic studies so far have done so, and [ii] our EDW-related results illustrate particularly well how this can help model assessment. We thus propose to explicitly cite 2 of these earlier papers, and to clarify (and shorten a bit) this paragraph as follows.

"Building upon a few earlier studies \citep[e.g.][]{Leroux2018, Fedele2021}, our present analysis illustrates the benefit of ensemble simulations over single hindcasts for model evaluation in the eddying regime. The random phase of CIV "noise" can result in either high, small or even negative model-observation correlations (from -0.45 to 0.8 for EDW temperature) depending on the ensemble member. Assessing a single eddying ocean simulation against observations should thus be done with care, all the more since observed fluctuations also contain random components, with an amplitude that will be specific to the object of study."