Response to reviewer comment RC1

Reviewer comments are presented first in *blue italics*, then followed by the author's response in normal font. Line numbers are referring to the original manuscript and are denoted as L145 for Line 145.

Major comments:

On the generation of the baseline: In this paper, the "baseline" state is created by re-running the model using perturbed initial conditions. However, deviations from reality do not arise only from uncertainty on initial conditions. Uncertainties on mixing parameterizations and meso-/submesoscale representation can also be important. Also, real observations always have more noise (internal waves, turbulence…). Random noise could be added to synthetic observations to increase their realism. By using a baseline generated by the same dynamical model, are you not placing your assimilation system in a very favourable situation?

We agree that deviations from reality do not only arise from uncertainty in initial conditions. They also arise from uncertainties in boundary conditions, atmospheric forcing, parameterisations and other model physics and unresolved physical processes (e.g. internal tides). However, the purpose of an OSSE is to test how effective the data assimilation of the chosen observation platforms is at improving the model estimates of the ocean state (given a known 'Reference State'). As a result, we assimilate into a model estimate of the ocean state that has errors that are typical of a 'real' operational DA system.To achieve this, we perturb the initial conditions with an offset in time to produce errors of a realistic level in the initially perturbed ocean state. By perturbing the initial conditions as we have done, we achieve a model estimate of the ocean state that deviates considerably from the Reference State in terms of the temporal and spatial evolution of the eddy field (as discussed in lines 150-155). We then assimilate synthetic observations from the Reference State that have random noise added. The magnitude of the random noise is such that the errors are normally distributed with a standard deviation equal to the realistic observation errors, as was explained at L179-181, L192-195, L207-2011. Our OSSEs are therefore showing how the DA minimises errors that come about from incorrect initial conditions (while the forcing, boundary conditions and model physics are the same). This systematic approach provides useful results to show the impact of surface and subsurface observations on the metrics that we evaluate.

We also discussed (L150-155) that we tested a variety of initial condition perturbations, and all led to errors growing to approximately equal amounts. Therefore, our choice of perturbation within the selection we tested is appropriate and justified (L147-149).

Lastly, as explained above, we do indeed add random noise to the synthetic observations (see explanation above). See L179-181, L192-195, L207-211 for more details.

Choice of synthetic observations: The authors have chosen to test XBT strategy but not any of the other observing systems, such as gliders, Argo, etc. Also, they chose to focus on temperature-only data, while salinity data are available on other platforms than XBT. Yet I imagine they should have an impact. Any reason for that? As it stands, this does not meet *the criteria of an objective assessment of the value and complementarity of different observing strategies?*

There are myriad observation platforms and strategies we could test. However, we purposely designed this study as a systematic approach by adding one datastream at a time. This approach allows us to explicitly test the value of each dataset as it is added to the system. This then explicitly shows the value of each XBT line separately and in concert, and allows testing of the upstream versus downstream impact of subsurface temperature observations. This was motivated by results of Kerry et al (2018) who showed observations of the eddy field were more impactful than observations of the upstream EAC jet, as discussed at lines 438-456.

Indeed in two earlier papers, we have assimilated glider and Argo data (see Kerry et al., 2016, and Kerry et al., 2018) to assess the impact of gliders and Argo data (including salinity) on the resulting DA simulation. Additionally, Argo and Glider data are more sparse in time and/or space, so it would not allow for a fair test of how they improve, for example, the subsurface structure of temperature. Whereas with a regularly repeated XBT line, we can robustly test this.

We have added the following sentences to the introduction to explain our reasoning for focussing on subsurface temperature observations, changing the following lines (L71-73):

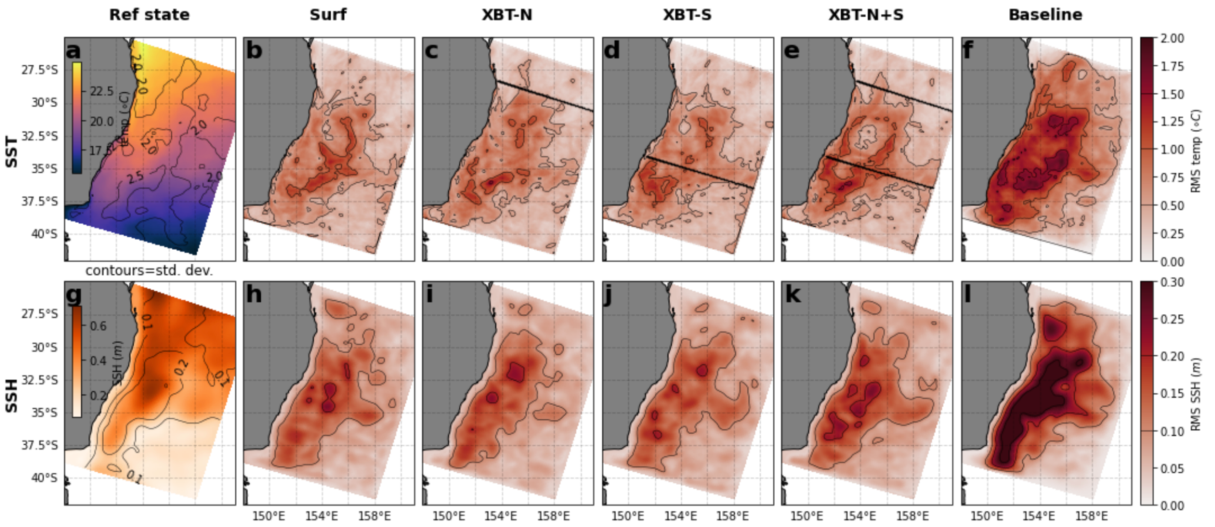
> “In particular, we examine the role of surface and subsurface temperature observations in improving the simulation of prominent EAC flow features, the vertical and spatial heat and velocity distributions, and ocean heat content.”

To

> “In particular, we have chosen to examine the role of SSH, SST and subsurface temperature observations in improving the simulation of prominent EAC flow features, the vertical and spatial heat and velocity distributions, and ocean heat content. Subsurface observations are systematically added in separate OSSEs to show the value of each observation platform in the absence or presence of the other subsurface observations.”

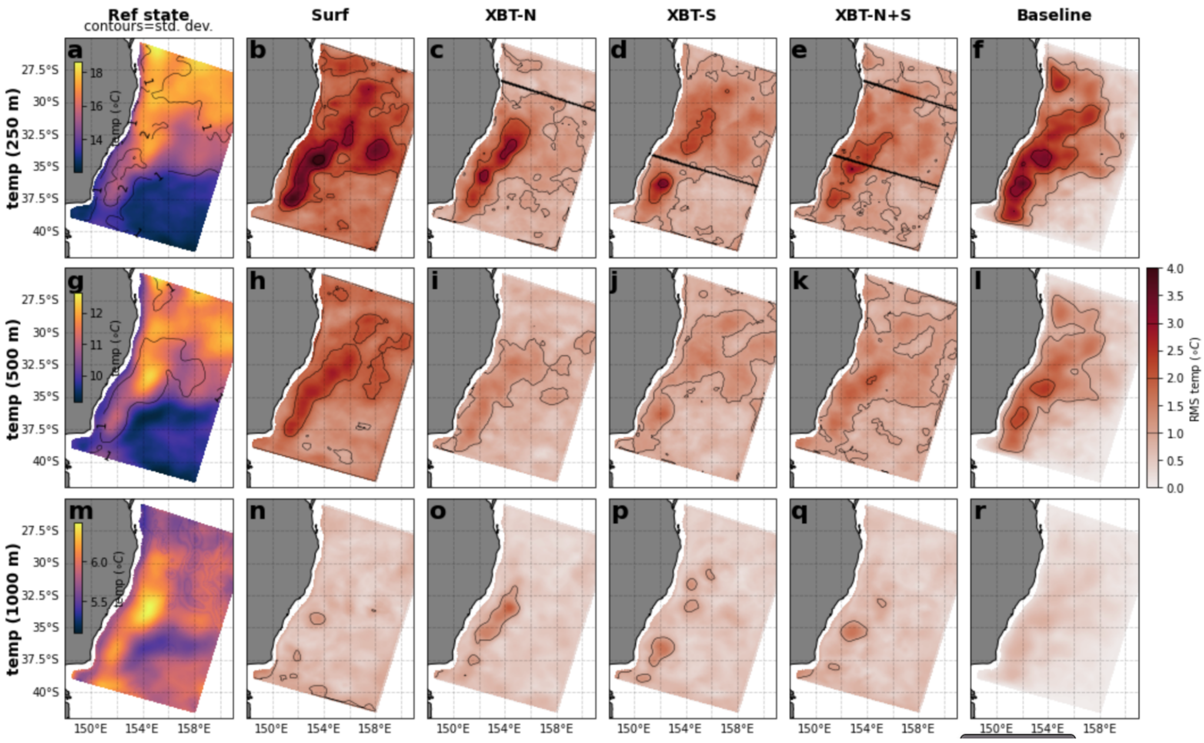
Conclusions are weakly supported by the results. The description of the beneficial impact of XBT observation seems exaggerated. In particular, it is very hard to assess the impact of different assimilated datasets as the baseline case is never presented anywhere. Another puzzling aspect is that the case assimilating all “data” is markedly less good than the one with only one XBT section.

We have followed the reviewers request and explored the comparison against the perturbed (‘Baseline’) free-running simulation. The plot below shows the SST and SSH fields in the Reference state, and the RMS error for that field in the Surf and XBT OSSEs, and the Baseline.



As expected, all OSSEs perform considerably better than the Baseline run at the surface, as assimilation of SSH and SST observations ensure the correct spatial and temporal evolution of the eddy field at the surface. This figure will be added to the supplementary (and referred to in the new paragraph (see below). Consequently, we have removed the reference to the SSH fields in Fig. B1.

We have also updated the original Figure 5 spatial RMS plot (see below) to include a comparison with the perturbed run. This demonstrates the improvement that the DA simulations provide over a free-running simulation, and is a useful addition to the manuscript. We thank the reviewer for this suggestion.



Despite the improved representation of the surface eddy field, the experiment that assimilates surface only observations (Surf) has greater errors than the baseline at 250m, 500m and 1000m depth. This highlights the poor subsurface representation when assimilating only surface observations, consistent with results of Zhang et al (2010) and Zavala-Garay et al (2012). All of the OSSEs that include XBTs (XBT-N, XBT-S and XBT-N+S) provided considerable improvement on Surf and provide errors that are below the Baseline run at 250m and 500m in the eddy field region. This is because the dynamic evolution of the eddy field is better simulated with data assimilation. RMS errors at 1000m are slightly greater for all OSSEs than the errors in the Baseline run, which is explained by a) the lower natural variability at this depth (necessarily making the free-running Baseline more accurate through lower impact of that initial perturbation) b) the corrections that the 4DVar simulation makes often leads to degradation in vertical representation.

Indeed, better subsurface representation of the ocean circulation, including eddies, requires further improvements to the 4DVar system configuration (specifically, improved background error covariance estimates) and is the focus of our future work. However, the purpose of this study is to show the value of subsurface temperature observations given a carefully configured ROMS 4DVar system (Kerry et al., 2016), with a particular focus on the EAC and EAC eddy field.

We appreciate the suggestion from the reviewer to consider the Baseline, and we feel it has raised some interesting points. So, in addition to updating Figure 5 to include the Baseline RMS fields, and the new supplementary figure of a SSH and SST comparison to the Baseline, we have added the following paragraph at line 290:

"Assimilation of SSH and SST observations ensures the correct spatial and temporal evolution of the eddy field at the surface, leading to all OSSEs performing better than the Baseline (cf. Fig. B2). All of the OSSEs that include subsurface observations produce better representation of subsurface conditions at 250m and 500m than the Baseline simulation (cf. Fig 5c-e to Fig 5f; Fig 5i-k to Fig 5l). This is because the spatial and temporal evolution of the dynamic eddy field is better simulated with data assimilation. RMS errors at 1000m are slightly greater for all OSSEs than the errors in the perturbed run (cf. Fig 5n-q and Fig 5r), which is explained by: the lower natural variability at this depth (necessarily making the free-running Baseline more accurate through lower impact of that initial perturbation); and, the corrections that the 4DVar simulation makes often lead to a degradation in vertical representation. From hereon, we exclude comparison to the Baseline so as to focus on the impact of the subsurface observations compared to the Surface-only experiment."

With regards to the impact of assimilating both XBT lines, as opposed to just one XBT line, Fig. 5 and Fig. 7 give examples of when assimilating all the observations (the SSH+SST+XBT-N+S OSSE) produces an overall better solution than when only one XBT line is assimilated. Temperature RMS at 250m is improved by XBT-N+S, compared to Surf, XBT-N and to a lesser extent XBT-S (Figure 5e compared to Figure 5b-d). Likewise, the vertical temperature RMS for the XBT-N+S OSSE across all transect locations, performs competitively compared to the Surf, XBT-N or XBT-S OSSEs (Figure 7). However, there are certain metrics for which this is not the case. To highlight this, we have modified the following line at L323:

"The XBT-N+S OSSE has relatively good representation of temperature along this transect and much improved representation below 1000 m (Fig. 7o)."

To:

“The XBT-N+S OSSE has relatively good representation of temperature along this transect and much improved representation below 1000 m (Fig. 7o); and indeed considering all transects (Fig. 7e,j,o), the XBT-N+S OSSE has low RMS error, as opposed to the single XBT OSSEs, which have high RMS at some transect locations (e.g. Fig. 7h,n).”

This is explained by the least squares best fit for both sets of observations being one with a degraded fit at both locations, as we discuss at L377-379. We have also added some further discussion of this at L422.

There are other examples in the literature where the addition of observations degrades estimates. For example, Zhang et al. (2010) showed that assimilating surface current observations degraded subsurface temperature estimates, and Siripitana et al. (2020) showed that assimilation of subsurface observations is often at the expense of the fit to surface observations.

Overall, I am not convinced this paper is best suited for publication in Geophysical Model Development. There is no model development presented here.

As per the journal scope, GMD does not only contain papers focussing on model development. This manuscript is submitted as an “model experiment description paper”. For the manuscript type, the GMD instructions for manuscript type 4, (https://www.geoscientific-model-development.net/about/manuscript_types.html#item4) note that “Configurations and overview results of individual models can also be included as well as descriptions of the methodology of experimental procedures such as ensemble generation. Such papers should include the discussion of why particular choices were made in the experiment design and sample model output.” Our manuscript is an overview of the results of a model experiment, and fits the criteria set out in the GMD manuscript types guidelines.

Minor comments:

1. *l. 1 and l. 16: WBCs are not necessarily poleward flowing. Maybe add here that you focus on Subtropical WBCs.*

We have modified L16 as suggested to:

“Subtropical Western Boundary Currents (WBCs) transport warm and saline towards the poles”...

The mention of poleward flow in the abstract is: “Western boundary currents (WBCs), form the narrow, fast-flowing poleward return flows of the great subtropical ocean gyres”, which is ostensibly correct.

2. *l. 23: Not all DA methods combine model-obs in a “dynamically consistent way”. State here that you are using 4DVar with 5 days windows.*

The text has been modified as follows:

Data Assimilation (DA) combines observations and a numerical model in a dynamically consistent way such that the result is a better estimate of the ocean circulation than either alone (Moore et al, 2019).

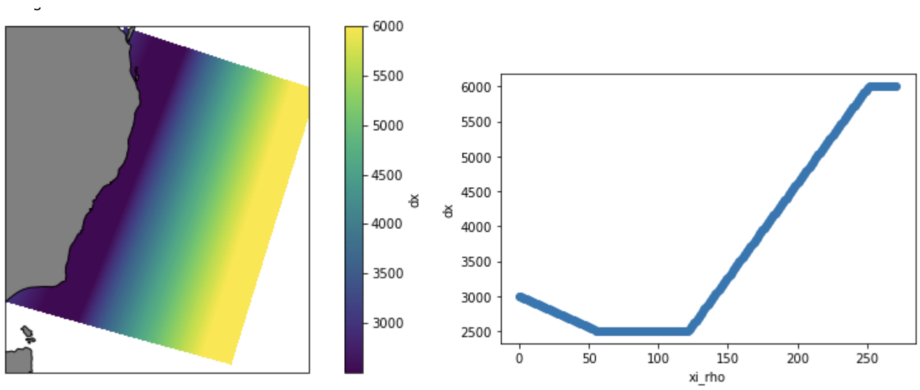
To:

Data Assimilation (DA) is the combination of observations and a numerical model, such that the result is an optimal estimate of the ocean state (Moore et al., 2019).

We state at L81 and L97 that we use 4DVar with a 5 day window.

3. *l. 86: why is the resolution varying so much in such a small domain? And how?*

The model resolution varies between 2.5 km on the continental shelf and slope to 6 km in the deep ocean over a distance of > 1500km, with a linear transition between. The gradual change in resolution is shown in the below plots:



These plots show the variation in horizontal resolution (in metres), with a gradual increase from 2500m to 6000m, across the domain in the across-shelf direction (xi_rho coordinate) of the model.

We have modified the manuscript to say:

"Horizontal resolution is 2.5km over the continental shelf and slope, and increasing linearly to 6km in the deep ocean"

There are also many previous studies that have used this model domain successfully (e.g. Kerry et al., 2016; 2018; Kerry and Roughan, 2020; Li et al., 2020; Li et al., 2021).

4. *l. 255: These results are very confusing. How is it possible that the worse experiment be the one assimilating the largest amount of data? Why do you not show the baseline case to see how bad the run is without assimilation? Looking at this figure only would actually lead me to conclude that XBT data have a very minor impact if not a detrimental one on the representation of the circulation.*

It has been widely shown that increasing the number of observations that the assimilated analysis has fit to will lead to a reduced fit to any one single observation. For example, Siripatana et al. (2020) found that including extra datastreams (HF radar currents and moorings) in addition to traditional observations (e.g. satellite SSH and SSH observations) degraded the model representation of SSH and SST. Zhang et al. (2010) showed that HF radar observations of currents degraded the subsurface temperature forecast, which they attributed to a lack of cross-variable covariance estimates.

We have demonstrated that subsurface observations improve estimates of key quantities (e.g. temperature at depth), which is especially noticeable in the high EKE region. It has also been widely shown that the assimilation of subsurface observations has a strong impact on ocean state estimates. For example, Moore et al. (2011) presented evidence for the considerable impact that subsurface observations have relative to surface observations. Zavala-Garay et al. (2012) also showed that surface-only observations result in poor estimates of the true subsurface ocean, and while this is improved dramatically in the presence of XBT observations, away from those observations, the improvement in subsurface skill was poor.

We have added the below paragraph to the discussion section at L421:

“We have demonstrated that subsurface observations improve estimates of key quantities (e.g. temperature at depth), which is especially noticeable in the high EKE region. It has also been widely shown that the assimilation of subsurface observations has a strong impact on ocean state estimates (Moore et al., 2011; Zavala-Garay et al., 2012). However, increasing the number of observations can lead to a reduced fit in any one single observation (e.g. compare RMS in XBT-N+S to XBT-S or XBT-N). This has been also demonstrated by others. For example, Siripatana et al. (2020) found that including extra datastreams (HF radar currents and moorings) in addition to traditional observations (e.g. satellite SSH and SSH observations) degraded the model representation of SSH and SST. Zhang et al. (2010) showed that HF radar observations of currents degraded the subsurface temperature forecast, which they attributed to a lack of cross-variable covariance estimates.”

We have addressed the comment on the comparison with the baseline in an above reply.

5. *l. 389: I do not understand which feature in the deep MKE is better represented thanks to XBT assimilation. On the contrary, it seems to me that deep properties are somehow worsened by XBT assimilation when compared with the Ref.*

We apologise for not being clear, we have replaced the discussion with the following line:

“The better reproduction of key EAC features in the surface MKE fields (e.g. more energy in the southern extension, return flow and eastern extensions; Fig. 4) when subsurface temperature observations (together with surface observations) are assimilated must be due to the improved subsurface conditions impacting the surface circulation.”

6. *Discussion is mostly a repetition of the Result section. What is the purpose?*

We have 5 sections in the discussion, each with their own sub headings (see Lines 382, 422, 437, 457, 492) each with a different theme. In this section we place our results in context of the literature and show how the different OSSEs resolve the dynamics of the EAC system and how the different experiments contribute to resolving upper ocean heat content in the system. The final section in the discussion is where we discuss the results in the context of observing system design - which is an important contribution to the ocean prediction and ocean observing community.