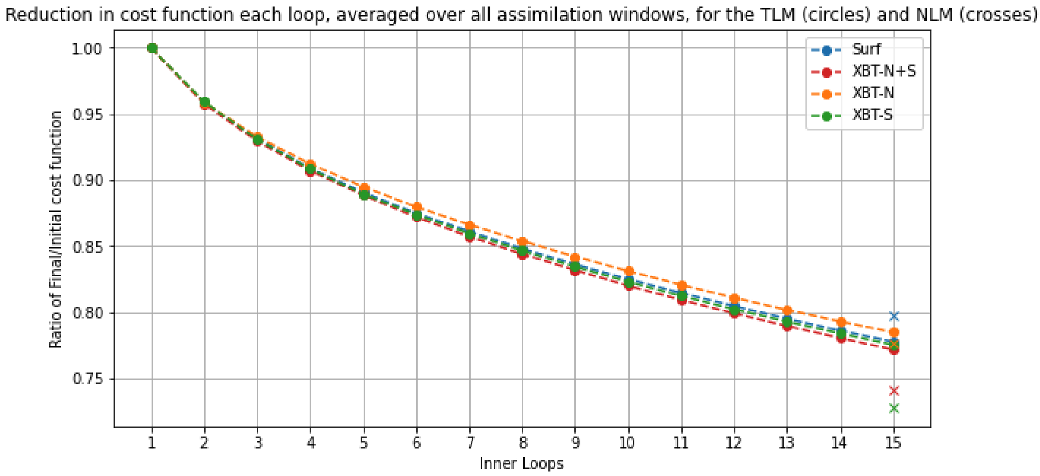
Response to reviewer comment RC2

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.

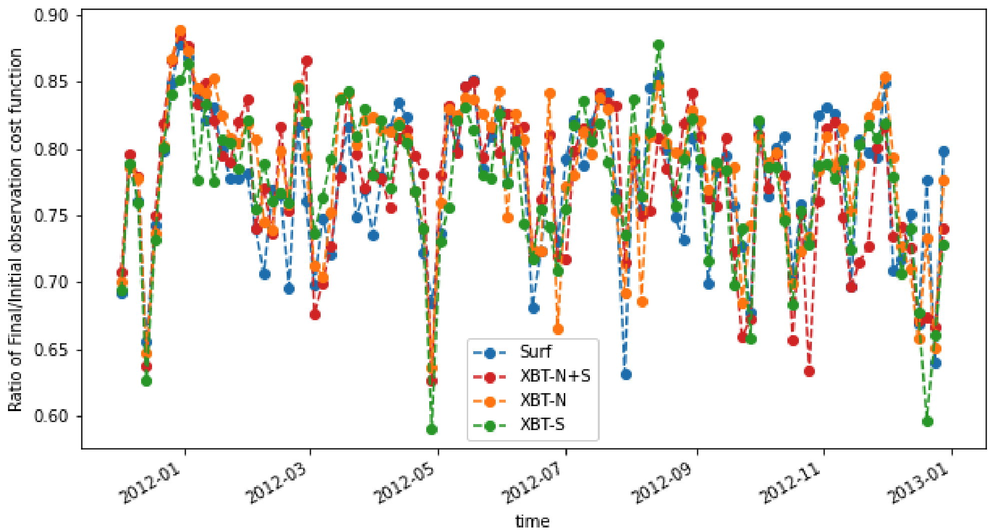
Major comments:

In general, if much observations are assimilated, the result will show higher accuracy. However, the authors have shown that the experiment assimilating both southern and northern XBT observations did not show the best performance. I am wondering if the cost function was properly reduced. How were the convergence conditions defined, such as the ratio of the final to the initial value of the cost function or the gradient of the cost function?

We chose to define a set number of inner loops to reduce the cost function. This was based on previous work of Kerry et al., (2016), who showed that 15 inner loops achieved an acceptable reduction with a reasonable computational cost, and is common in 4D-Var applications. The number of inner loops we used (15) shows similar convergence between all the OSSEs, with a likewise similar final nonlinear model cost function ratio (crosses). We show this in the below plot.



The ratio of the final observation cost function to the initial observation cost function shows minimal differences between the Surf, XBT-N, XBT-S and XBT-N+S OSSEs.



This shows that all OSSEs have similar cost function reduction, with the mean values of the ratio of final to initial total cost function ranging between 0.73 and 0.80 and the time-mean ratio of final to initial observation cost function ranging between 0.77 to 0.78. That is, all OSSEs have similarly reduced cost functions. Please refer to Eq 6 in Kerry et al, 2016) for more details of the cost function.

We now include in the text the following sentence at L98:
"We chose 15 inner loops to reduce the cost function, based on previous work of Kerry et al., (2016), who showed that this number of inner loops achieved an acceptable reduction with a reasonable computational cost. Similar cost function reduction is achieved for all OSSEs, with a time-mean ratio of final to initial cost function ranging between 0.73 to 0.80."

And following from that, we added the following sentences to elaborate on the quality control methods we employed:
"Background quality control was applied to eliminate observations that are poorly represented, following the method described in Moore et al (2013). Only observations that satisfy $d_i^2 < \alpha^2(\sigma_b^2 + \sigma_o^2)$ are assimilated, where d_i is the innovation, the quality control parameter α=4 and σ_b and σ_o represent the prior background and observation errors. In all OSSEs, about 20% of SST observations were rejected by this criteria while all of the SSH observations were assimilated. For the subsurface XBT observations, all observations were used for the single transect OSSEs, while in XBT-N+S between 20-40% of observations were rejected, due to the innovation being too large. This method has been applied in several other recent 4D-Var studies (Levin et al., 2021; He et al., 2022)."

The authors also show the XBT-S experiment is the best among the OSSEs. How does the observed information propagate upstream (to the north)? As I see the Figure 6 in Kerry et al. (2018), the impact of the XBT is limited to the vicinity of the observation latitude. An influence scale of 600 km seems distant relative to the spatial scale of analysis increment. If there is an impact of the XBT-S observations on the northern boundary, it should be clarified in the manuscript.

There may be some misunderstanding of Figure 6 of Kerry et al,. (2018): In that figure, the XBT observations taken between 34S to 36S have an impact on transport at all cross sections between 27.5S and 36.2S. Further, although the total impact is low, the impact per observation (bottom panels) is high.

We are unsure what the reviewer means by "An influence scale of 600 km seems distant relative to the spatial scale of analysis increment." Analysis increments are applied to the boundary, surface forcings, and initial conditions, with the spatial scales controlled by the background error covariances as described in Kerry et al. 2016. The use of the adjoint and tangent linear models in 4D-Var introduce flow dependence to the covariances.

It is true that the information provided by the observations cannot propagate 600km upstream by advection over the 5 day windows, but in 4D-Var, information can propagate via a number of physical processes as discussed in Kerry et al, (2018). In that study, observation impact is shown to be far reaching both up and downstream. There are other examples of far-reaching observation impact in the literature, for example in Siripitana et al (2020) and Powell et al., (2017).

These points are discussed in section 4, but it would be better if you could briefly introduce the discussion part in the results section, for example after L316. When I read section 3.4, I wondered how the southern XBT line would affect the upstream regions far from 600 km.

In this paper we have tried to mostly limit the results section to focussing on results, and the discussion section to the discussion of the key ideas emerging from the results. However, we take the reviewers point, and have added at L316:
"The mechanisms that might lead to such action at a distance are discussed further in Section 4.1 and 4.3."

Likewise, we have added further discussion of the results showing the reduced impact from XBT-N+S at L323:

"The XBT-N+S OSSE has good representation of temperature along this transect and much improved representation below 1000m (Fig.7o; depth-averaged RMS of 0.6C); 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).

And more at L422:

"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."

The last two additions were in response to RC1, and we refer to that answer for further details.

This study focused on the XBT observation network. However, in realistic situations, other networks exist, such as Argo floats, HF radars, and sea gliders. Why did the authors investigate the impact without using other observation network? Of course, I admit that it makes sense to evaluate the impact of XBT observations.

As RC#1 also made this remark, we answer similarly: There are many 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.

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.

Following RC#1 and this review, 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."

Temperature observations per model cell may be too dense to account for representation errors. It may be better to consider the super-observation or thin out the observation. Also, if you want to show the advantage of the high density XBT observations, it would be useful to demonstrate an additional experiment which assimilates the XBT at regular observation interval. In addition, the manuscript mentions observation errors of XBT, but it is unclear whether the representation errors are considered. How about you clarify this point?

We only apply a single observation per cell. Further, as the synthetic observations come from the same model grid, there is no need to generate super-observations as only one data point per cell is extracted.

We agree that exploring the impact of the XBT transect density would be useful and interesting. However, to properly explore this, it would require more model runs and enough additional analysis that it would broaden the scope of this manuscript too much.

Representation errors result from the discretisation of the model grid and unresolved physical processes (e.g. inertial tides). However, none of these sources are relevant to us as we sample the observations from a free-running model that has the same model grid and resolves the same physical processes. The errors we apply add noise such that the observations are realistic of real observations (where there would be representation errors).

We have added the modified sentence to clarify this point at L185, from:
“As a result, we choose each model point as an observation location.”
To
“As a result, we choose each model point as an observation location, and have no need to superpose or thin observations. Further, and likewise for the other synthetic observation types, there are no representation errors, as we sample observations from the same model; realistic errors are added with random noise (see below).”

Also note that we compare the realistic XBT transect temporal and spatial resolution (L199) to our synthetic XBT transect temporal and spatial resolution (L200 & L203).

The authors often use the term “best” to specify which experiment represents good performance (such as L262 and L356). However, it should be better to quantify the experiment using statistical metrics such as area-averaged biases and RMSEs. In addition, it would be useful to specify the improved ratio of each OSSE relative to the surf-only experiment or the baseline when comparing the impact of observations across each OSSE (for example L297 and Figure 6).

In light of the reviewer’s suggestion, we have added the area-averaged RMS error value to the corner of each panel in Figures 5, 6, 7, 10 and the new Supplementary Figure B2. Because the eddy region (box b, as shown in Figure 3b) is the region of highest RMS and is of particular interest for representing eddy dynamics, we have also included the mean RMS for this region, and both values are noted in each panel. We have updated the captions accordingly.

We have adjusted the text to include better quantification with these RMS values:

At L274:
“For each field and OSSE, the area-mean RMS value, as well as the mean RMS error for the high EKE region (box b; Fig.3b) are shown in Fig.5.”

At L278:
“The presence of the subsurface observations greatly improves 250m temperature RMS (Fig.5c-e; compare RMSE values in the eddy region of 1.6C - 1.8C to 2.5C for Surf).”

L282:
“The addition of the north and south transects in the XBT-N+S OSSE (Fig.5e) improves temperature RMS compared to the Surf OSSE, and has a similar spatial pattern in RMS error to the southern XBT OSSE, with a mean RMSE of 1.7C in the eddy region, compared to 2.5C for the Surf OSSE.”

L285:
“The OSSEs with either a northern or southern transect of XBT observations (Fig.5i-j) display relatively low RMS error, especially compared to the surface only observations (Fig.5h; area mean RMS of 0.9C compared to 1.7C for Surf).”

We have added a line to the discussion on the Baseline simulation (see RC#1 comments):

“The area mean RMS values for the Baseline are marginally lower than XBT-S for temperature at 250m and all OSSEs at 500m; when considering the mean over the eddy region, the Baseline simulation has a higher mean RMS than the subsurface OSSEs. This is because the spatial and temporal evolution of the dynamic eddy field is better captured with data assimilation.”

L319:
“The addition of subsurface observations in either the north or the south reduces RMS error near the separation zone (Fig.7h-i; depth averaged RMS values of 0.7C, as compared to 1.0C for the Surf OSSE), indicating an improvement in the representation of heat content carried by eddies.”

L321:
“The XBT-N+S OSSE has good representation of temperature along this transect and much improved representation below 1000m (Fig.7o; depth-averaged RMS of 0.6C)...”

L368:
“...however observations in this upstream region (Fig.10b,d; XBT-N and XBT-N+S) lead to the largest improvement in error in representation of the EAC jet, with depth-averaged RMS values of between 0.08m/s -- 0.09m/s, as compared to 0.1m/s for the Surf and XBT-S OSSEs.”

L372:
“During northern separation at ~32S (Fig.10i), both experiments with a single subsurface XBT transect (XBT-N and XBT-S, Fig.10j-k) have the lowest RMS error (depth averaged RMS of 0.14m/s compared to 0.19m/s for XBT-N+S and 0.25m/s for Surf),...”

L376:
“...velocity representation along the ~32S transect is better in the presence of southern rather than northern subsurface observations (compare Fig.10o to Fig.10n; RMS error of 0.08m/s compared to 0.13m/s).”

With regards to the reviewer's second point, we originally calculated the ratio of improvement, defined as the percentage ratio of the RMS of the XBT-N, XBT-S and XBT-N+S OSSEs, compared to the RMS of the Surf OSSE. However, we went back to pure RMS values, because it was easier to interpret the figures and associated particular patterns in the plots with physical mechanisms, for example, increased RMS in the Surf OSSE in the subsurface velocities near to the return flow.

The upper ocean heat content (OHC) in the OSSEs in Figure 8a and 8b showed similar temporal evolution until March or April 2012. Why the observation impacts were less? Conversely, why did the difference of the upper OHC occur after April 2012? Does it relate to the EAC phase?

We thank the reviewer for this comment. It seems likely that the patterns described do relate to periods of southern and northern EAC separation. For example, the OSSEs do show similar evolution in UOHC until around March, before they diverge in their temporal evolution. This coincides with a period of southern separation (until an eddy shedding event in mid March). Conversely, in late September, the EAC separation region moves northwards, which coincides with a period of increased difference between the OSSE predicted UOHC and the Ref state (e.g. see Figure 8a).

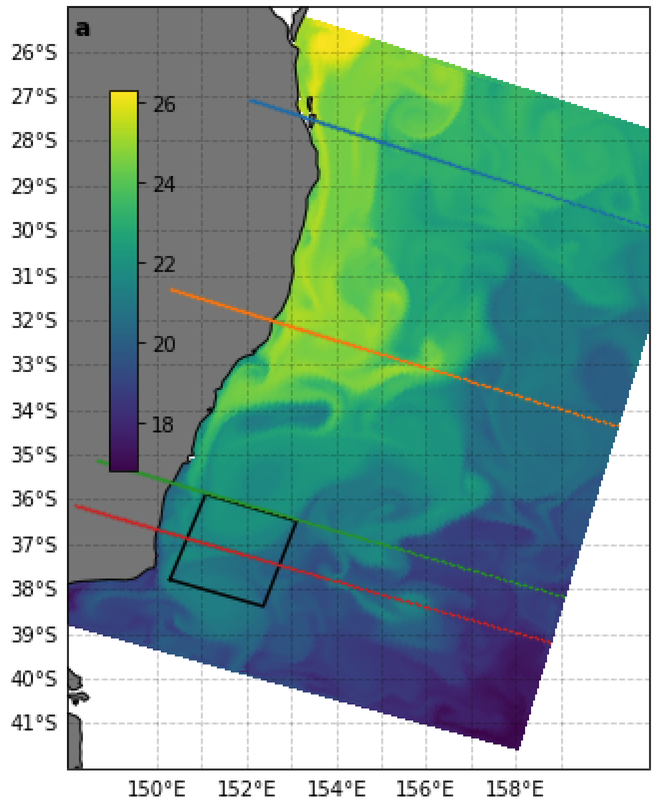
We have added a new paragraph to highlight this, at L342:
“The periods of increased or decreased error between the Ref state and the OSSEs, as well as between each OSSEs, generally coincide with periods of more northerly or southerly EAC separation latitude (See Section 3.6 and Section 4.4). For example, during mid January and May 2012 there was relatively good agreement in UOHC, which coincided with periods of southern EAC separation. Conversely, in October, when the EAC separated further north, most OSSEs had increased error compared to the Ref state. Interestingly, while all OSSEs have generally worse representation of UOHC in the upstream box a region, there are periods of lower RMS, for example in early to mid August, when there is a coherent EAC jet through the box a region, and likewise, with box c in late April. This suggests that the ability to represent UOHC depends on the location of the coherent EAC jet and separation latitude.”

This now nicely segues into the next section on separation latitude. We have added an additional note at L471:
“The error in UOHC between the OSSEs and the Ref state is also related to separation latitude, as discussed above (Section 3.5).”

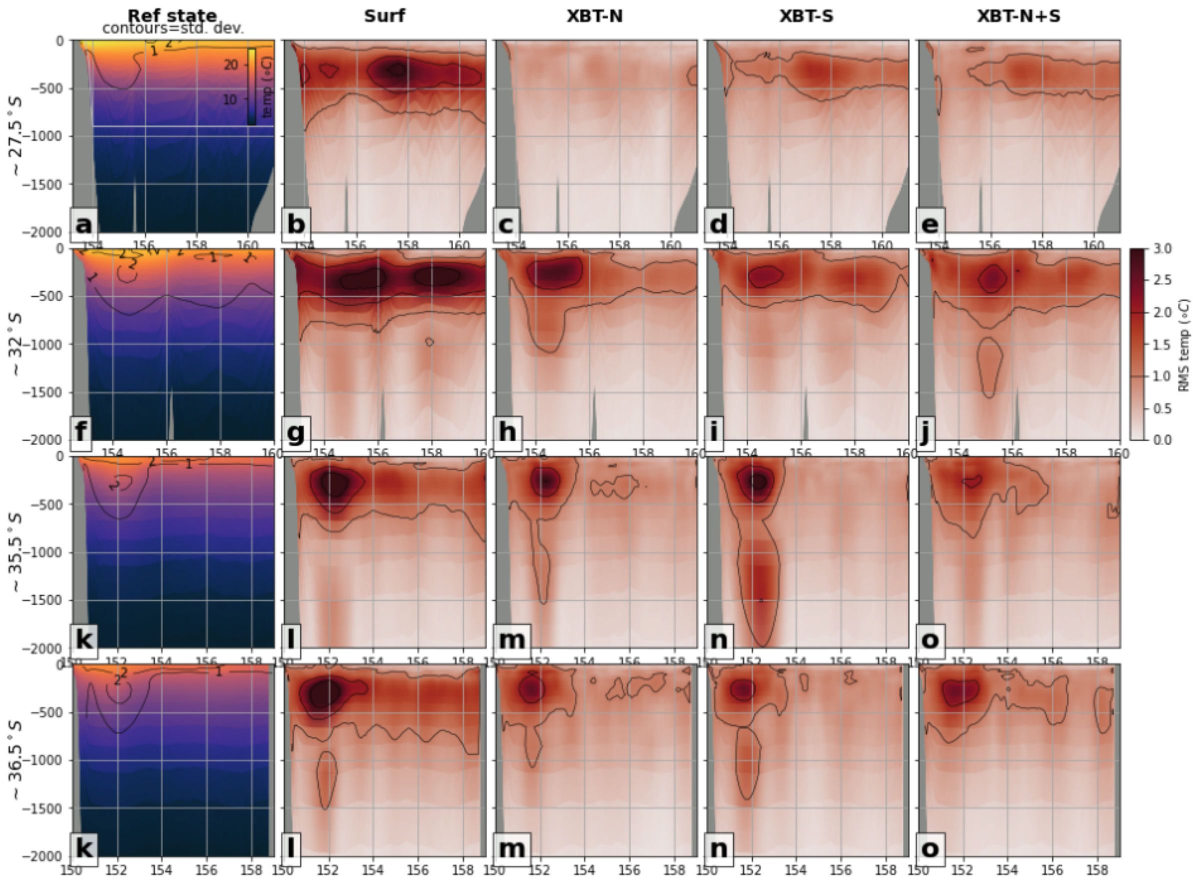
The vertical temperature section for the ~35.5S transect in Fig. 7n shows a large RMSE below 1000 m depth, whereas the upper OHC above 2000 m of the XBT-S in Fig. 8b and 8c (green dash line) is better represented than in other experiments. Based on the fact that the 35.5S transect locates between box b and box c shown in Figure 3b, the vertical temperature in Figure 7n would be expected to be best represented. However, it is not, and appears to be inconsistent. How should we interpret this point?

This apparent inconsistency is because the UOHC in box c is further south than the transect location of 35.5S.

To demonstrate this, we present the transect through the middle of box c, as demonstrated here (red line is at ~36.6S):



The mean RMS in temperature through this transect is shown in the below plot:



Now we see that the RMS error is lower at depth in XBT-S (compare panels s to n), as well as lower RMS in XBT-N & XBT-S compared to Surf. This matches what we see in the UOHC time series plot, demonstrating there is no inconsistency.

Specific comments

P5 L134: The Reference state and the baseline experiment is forced by BARRA-R while the OSSEs have applied the ACCESS reanalysis as surface forcing. Is it right? If so, how about you emphasize that such condition leads to additional perturbations?

That is correct. We have added the following sentence to L155:
"Note that the different surface forcing conditions between the free run (BARRA-R) and the OSSEs (ACCESS) leads to an additional source of error that the DA system must reduce."

P6 L143-149: Was the initial condition for OSSE chosen from the Reference state 8 days later or 8 days earlier as perturbation? In other words, is it a lagged initial condition? It would be easier to understand if you describe the date of the initial condition used in OSSE.

We have modified this line as follows:
"We initialised each OSSE with initial conditions that were 8-days offset from those that were used to initialise the Reference state."
To:
"We initialised each OSSE with initial conditions that were 8-days offset from those that were used to initialise the Reference state (i.e. begin the OSSE at 02 December 2011 with conditions from 10 December 2011)."

P9 L198: Transect PX30 represents the Brisbane to Noumea line (southeast direction) in this study. However, in the previous studies (Kerry et al., 2016, 2018), PX30 refers to the Brisbane - Fiji route (northeast direction). Does this mean the observation line PX30 has changed?

We thank the reviewer for pointing this out. As seen on the NOAA website, the PX30 line goes from Brisbane to Noumea or Fiji (http://www-hrx.ucsd.edu/px31.html). We also make clear that our XBT transects are chosen to approximately represent these ship tracks: "XBT locations were generated to very approximately match XBT deployments.." at L197.

P20 L398: It should be better mention about the EKE based on the Figure B.

We have added a reference to the figure at the suggested location, and in the following sentence:
"Surface EKE is not further improved with the assimilation of temperature observations at depth (see Fig.B2a-e). Furthermore, the improvement from temperature (XBT) observations on EKE at depth (500m) is also minimal (Fig.B2g-j)."

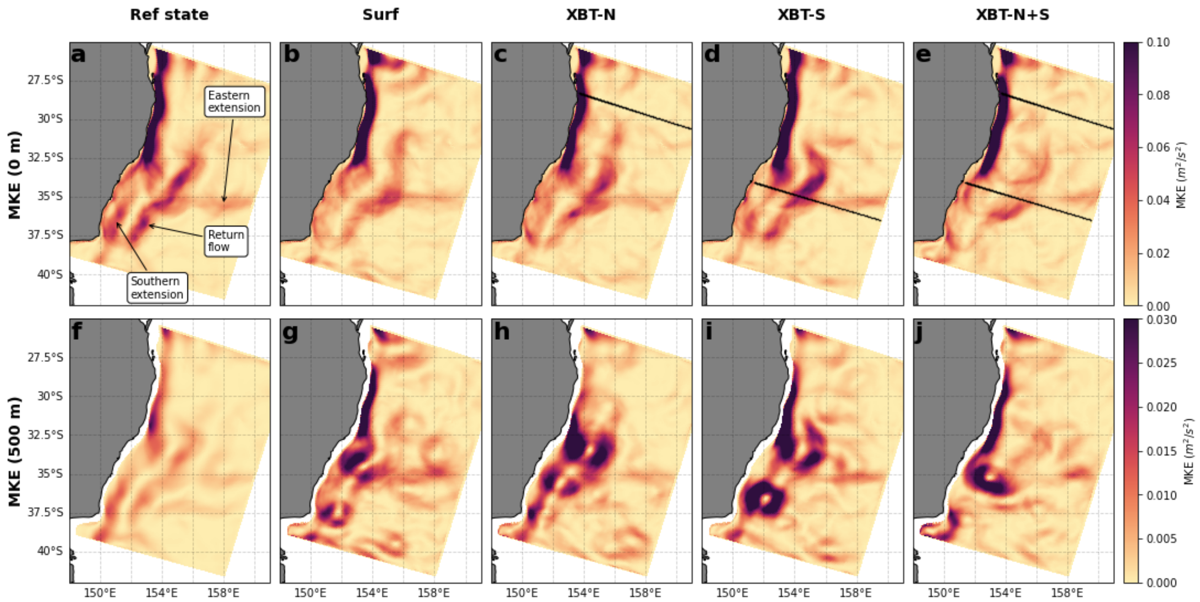
Technical comments

P4 L82: Please add "variational" for the abbreviation of "4DVar".

Thank you - added.

P11 Figure4: It would be useful to point out the velocity directions for the EAC southern extension, the EAC return flow and so on. It is also in Figure. 5 and 9. This study will attract not only the oceanographers but also data assimilation community. It will help for the readers who are not familiar with the EAC circulation.

We have updated Figure 4a to include labels to the EAC Southern Extension, Return flow, EAC Eastern Extension, as shown below:



The caption has been updated to reflect this, as well as to point out the flow directions of these long-term features:
"Key circulation features that emerge in the long-term mean are shown, including the EAC eastern and southern extensions and the return flow, which flow eastwards, southwards and northwards, respectively."

P16 Figure 7: Please draw the temperature color-bar for the Ref state in an easily recognizable location.

We have moved the temperature colour bar as suggested.

Other changes

Data repositories: Following the comment from the Chief Editor (see CEC#1), we updated our data and code availability statement (see CEC#1 reply):

"The source code, forcing conditions, configuration files and output for the simulations conducted here are available at https://doi.org/10.26190/unsworks/24146. The free-running EAC ROMS model forcing conditions are sourced from the Commonwealth Science and Industrial Research Organisation (BRAN2020; available at https://research.csiro.au/bluelink/outputs/data-access/) and the Bureau of Meteorology (BARRA-R and ACCESS; http://www.bom.gov.au/research/projects/). Along-track SSH data is available from the E.U. Copernicus Marine Service Information (https://doi.org/10.48670/moi-00146). Model configurations for the free-running and DA simulations are identical to those used in previous simulations (available online at https://doi.org/10.26190/TT1Q-NP46; https://doi.org/10.26190/5ebe1f389dd87). The model source code is open-source and available from https://www.myroms.org/."

We have added several sentences at L493 to mention the use of the Surf-only DA setup for SAR:

"Firstly, the presence of surface observations (SSH and SST) significantly improves the representation of surface currents, which is most important for navigation and search-and-rescue applications. "

At L91, added "are" between "conditions" and "given".

Figure 3 caption: removed "with with".

At L170, replaced "mission" with "missions".

At L276-277, added "(compare to the Baseline; Fig.5f)" to make clear that Surf degrades representation compared to the baseline.

At L292, replaced "of a subsurface" with "of subsurface".

At L328, replaced "is shown" with "are shown".

At L473, replaced "separation is" with "separation are".

At L479, replaced "phase is" with "phase are".

At L486, added "in recent decades".