Response to Reviewer 2

We would like to thank K. Andrew Peterson for his thorough review, which will help to improve the overall quality of the paper. Responses to the reviewer's comments are highlighted in blue.

Major Comments

A large portion of the justification for removing the longer covariance length scale from the balanced portion of temperature background covariances – and therefore sea surface height and salinity increments – is based on the balance operators used in the NEMOVAR assimilation scheme. I believe a more detailed reminder of what those balance operators are beyond the stated "physically-based balance relationship" (l. 146) is required. The balance relationships of dynamical height adjustment and maintenance of neutral density, are critical to understanding why the longer temperature covariance length scale may be sub-optimal, but the contribution to an unbalanced salinity increment might have lesser negative effects. Although this might be obvious to the authors, it is sometimes necessary to lower the conversation to the readers potential ignorance. While the Waters et al. [2015] reference (BTW: An addition of the DOI to the reference would be useful: https://doi.org/10.1002/qj.2388) does extensively discuss this, I believe a more explicitly stated reminder would be useful here, possibly further referencing some of the earlier linearization work. Particular emphasis should be placed on why the introduction of a large scale covariance on temperature might bleed corrections to the unbalanced, barotropic sea surface anomaly (SLA) into the sub-surface temperature. By the way, the statement (l. 188) that "large-scale error covariances have a 400 km length-scale for temperature, unbalanced salinity and unbalanced SSH" is correct [Mirouze et al., 2016], but it is also confusing! With the start of that sentence discussing the shorter length scale covariances, this statement somehow suggests to me that unbalanced SSH also has two covariances length scales, when in reality, it only has the longer, 400 km length scale.

Our initial intention in the first paragraph of Section 3.1 was to give a brief context of the NEMOVAR balance operators, without giving too much detail about them, as we wanted to focus on the subsequent impacts of the DA changes. Therefore, we thought that having a proper reference, such as Waters et al. (2015), would be enough since the balance operators are properly described there. However, we agree with the reviewer that some of the GOSI9 DA impacts are related to the balance operators. Therefore, to accommodate the reviewer's comment without disrupting the flow of the paper around the new DA changes, we wrote a few more sentences about the NEMOVAR balance operators and cited previous works in addition to Waters et al. (2015). The introductory content of NEMOVAR and the description of the balance operators have been moved to the beginning of Section 3, so that Sections 3.1 and 3.2 are solely focussed on the impacts of the GOSI9 DA changes.

Regarding the confusing statement in Line 188, we have split it into two sentences. The first one is related to the short length-scales whereas the second is related to the long length-scales, so hopefully this will avoid ambiguities.

I believe Figure S4 should be elevated to the main article. Discussion of why the SSH balance relationship is now applied throughout the mixed layer in GOSI9 (or why it was turned off through the mixed layer in GO6), is quite brief, and seems quite conclusively demonstrated by Figure S4. Furthermore, it is equally, if not more, important, than the subsequent discussion of then mitigating some of the negative effects of that change (along with the removal of the long temperature covariance lengthscale) via the Brunt-Vaisala criteria in subsection 3.2. What is not expanded on, however, is whether the change to applying the balance throughout the mixed layer is related to the change to a single temperature covariance length scale. (I.e. I presume previous versions of the system better performed when this relationship was not applied in the mixed layer.)

Both the Brunt-Vaisala criteria and the changes in the SSH balance are applied to mitigate localised numerical instabilities exacerbated by DA when the long temperature covariance length-scale is removed. We changed the text to make this point clearer in Section 4.1:

"This change reduces DA water column instabilities in areas of water mass convection, such as the Mediterranean Outflow region**, which are exacerbated by removing the long length-scale component of temperature background errors (see Section 3.2)**. Applying the SSH balance throughout the whole water column results in improved SLA statistics in the Mediterranean Outflow, particularly for ORCA12 (see Fig. S4)."

Regarding Section 3.2, we have made a mistake in Fig. 4d, selecting a much smaller area than the Labrador Sea to calculate the observation-minus-background (OmB) statistics. This has been fixed and Fig. 4d now shows a much larger, convincing impact of using the Brunt-Vaisala criteria to reject T/S increments in the Labrador Sea. As both Fig. 4 and Fig. S4 are mitigating the negative effects of not having the long temperature covariance length-scale, we prefer to keep Fig. 4 in the main article and leave Fig. S4 in the supplementary material. In our opinion, there is no need to have both in the main article.

I found the rationale for applying the Brunt-Vaisala criteria to determine if the increment should be retained as less than convincing, at least with the current discussion and presentation of Figure 4. Showing the change of mean state at 1200m is a little confusing, as judging from the mean error shown in the profile of Figure 4d, it is not immediatedly obvious that the mean state of GOSI9+BV at that level is better than GO6 – and it is only just improving from being worse even then GOSI9 without BV slightly higher in the water column. I would think it might be better to highlight the mean temperature somewhere around 500m, where you can show that a more accurate mean state is achieved by not retaining the increment – and thus possibly increasing slightly the RMSE, although this latter part of the argument depends somewhat on whether the OMB statistics displayed are from the IAU phase (assumed), or the trial phase. This then can also lead, by not instigating spurious deep convection, to better RMSE and mean error lower down (below 1000 and 1300m respectively) in the column. I do not believe it is completely accurate to describe the mean state as being improved from 250m and below (l. 247), as it is fairly obviously degraded at 750m, at least relative to GO6. Finally, the Figure S6 is perhaps more definite in showing the improvement – although here the cost (increased RMSE) of not retaining the increment at certain levels is not as obvious. I would keep with the current discussion of the improvement, when well described, in the Labrador Sea deep convection region, but note to the reader further evidence for the Mediterranean overflow region is shown in the supplementary material.

As mentioned previously, the Labrador Sea area used for selecting the T profile observations in the OmB statistics was much smaller than it should be. Fixing this improves considerably the results of the OmB temperature statistics in the Labrador Sea, as shown by the corrected Fig 4d. The benefits of using the Brunt-Vaisala criteria are now clear in the Labrador Sea, with consistent impacts in the 500-2000 m mean differences and RMSDs. This is also consistent with showing temperatures at 1200 m in Fig. 4a-c, since the impacts of applying the Brunt-Vaisala criteria are clearly larger at depth in the corrected Fig. 4d. Fig. 4a-c also had their colourmap range decreased, making it easier to see temperature differences at 1200 m between the experiments.

The OmB statistics are from a one-day model forecast before the increments for that cycle have been included. In Section 4.1, we believe the statement below clarifies that:

"Although these observations are compared to the model background, **i.e. before being assimilated**, they cannot be treated as fully independent datasets".

Finally, I found the discussion of GOSI9 improving the overturning circulation (AMOC) in the latter portions of Section 4.3 (ll. 493 – 509) also not particularly convincing, at least by how it is served by Figure 14. The principle conclusion of section 4.3 in general, is that GOSI9 is more stable to the lack of sub-surface observations, or more specifically, mitigates the negative consequences of sea level anomaly assimilation in the absence of correcting sub-surface observations. The discussion on the AMOC attempts to then extend this improvement when lacking sub-surface profiles to the AMOC. This is convincingly demonstrated for the AMOC at 26.6◦N (Figure 14a), where the divergence of the non-profile assimilation run with regards to the underlaying RAPID observations is convincingly shown. This is less convincingly shown at 50◦N (Figure 14b) in the absence of underlying observations, which could be rectified by instead considering overturning along the OSNAP section, for which observations do exist, and the FOAM system has previously been evaluated [Jackson et al., 2019]. This is even less convincingly demonstrated by the the AMOC at 30◦S, where the GOSI9, no T/S profile assimilation (cyan) line appears to be the outlier instead of GO6 no profile (orange), and appears to be going in the unwanted direction of too small, and even negative southern hemisphere AMOC that resulted in FOAM not being used in the Mignac et al. [2018] analysis. I would encourage the authors to at least include an AMOC estimate along the OSNAP section as previously shown in Jackson et al. [2019], with accompanying observational estimates, which should exist for the examined 2019

We have now included the comparison between the model and the OSNAP transports in Fig. 14b. It is very consistent with the comparison with the RAPID transports, showing that GO6-NoTSProf is clearly overestimating the transports along these two arrays, while GOSI9-NoTSProf follows very closely the observed transports. This now demonstrates more convincingly the GOSI9 impacts on the ocean circulation, especially in the North Atlantic.

We agree with the reviewer that GOSI9-NoTSProf shows weaker transports at 30S in June-October 2019, but they seem to recover in November and December, being similar to the other runs. This is why we state in the paper that a long reanalysis should be run in the future, so we will be able to conclusively evaluate the GOSI9 impacts on the South Atlantic MOC.

Furthermore, I disagree with the implicit statement – it is not actually stated, but just somewhat implied by the discussion – that GOSI9 might improve (ll. 552-556) the "discrepant" equatorial overturning FOAM demonstrated in GO6, which is only somewhat, and not conclusively, demonstrated through Figures 14d-h, or the "unrealistic" southern hemisphere AMOC suggested by Mignac et al. [2018], which is not demonstrated at all in Figures 14d-h as the figure does not seem to go as far south 2 as 35◦S in order to compare with Mignac et al. [2018]. Although as stated (ll. 501-503) some of this might await a proper evaluation of a FOAM reanalysis.

In Fig. 2 of Jackson et al. (2019), the AMOC stream function of the Met Office reanalysis is compared to several other reanalyses. It is clear that the Met Office reanalysis has a much stronger equatorial transport, including discontinuities in the AMOC stream function, when compared to other reanalyses. This is a very similar pattern to what has been shown in GO6-NoTSProf with too strong equatorial transports (above 40 Sv) and AMOC discontinuities near the equator. Both aspects have been improved in GOSI9-NoTSProf, with AMOC transports decreasing by ~15 Sv in the equatorial region relative to GO6-NoTsProf, which is a clear improvement in the equatorial transports. This has been added to the paper when discussing Fig. 14.

Mignac et al. (2018) and Jackson et al. (2019) raised similar concerns about the South Atlantic transports in the Met Office reanalysis. Jackson et al. (2019) stated the following:

"In particular, GloSea5 is suspect in the South Atlantic and near the equator (where there is a discontinuity in stream function strength): This issue has been traced to the method of assimilating sea surface height and will be the subject of a future publication".

Although we agree with the reviewer that South Atlantic MOC improvements are still inconclusive, this paper addresses the concerns raised by Jackson et al. (2019). These concerns are related to how our sea surface height assimilation impacts the ocean circulation, which has been improved in GOSI9. A GOSI9 reanalysis from 1993 to present is currently being run and a future publication will be able to show conclusive results in the South Atlantic circulation.

Minor Comments

The manuscript mentions several assimilation method and ancillary input changes (ll. 71-74), along with the testing of new observational types (ll. 72-73). Most of the assimilation method and ancillary information changes have been implemented (or continue to be used) in GOSI9, but none of the mentioned new observation types are used – for the most part, due to the fact that they would not be operationally available. It might be interesting to return to this in the summary (i.e. could SSS assimilation [Martin et al., 2019] further allow removal of the longer salinity covariance scale; could U/V assimilation [Waters et al., 2024] improve equatorial transports, etc.).

This is a good point. We have added the following sentence in the summary:

"Additionally, the potential of assimilating new observations in future operational versions of FOAM, such as the satellite sea surface salinity (Martin et al., 2019) and surface current velocities (Waters et al., 2024), may further improve near-surface salinity fields and equatorial transports in GOSI9, respectively."

All Figures: It is not mentioned whether the observation minus background statistics shown throughout the manuscript come from the trial run or from the IAU run. It is known FOAM calculates both, and it would be useful to know which of the two is shown throughout the manuscript – assuming it is not a mix of the two. This would help with the interpretation of figures such as Figure 4; it would be natural for RMSE of the IAU step to increase at levels where the increment is not being applied.

This has already been addressed in a previous comment. However, we reinforce here that all OmB statistics shown in the paper are from a one-day model forecast before the increments for that cycle have been included.

(ll. 230-232) The adjustment of melt pond variables to changes in SIC DA changes seems quite important to improvements/changes in sea ice for GOSI9 (ll. 393-403). It is also a rather ambiguous statement. Further details would be warranted/appreciated.

The increments are added to the sea ice concentration on each IAU step, and the sea ice volume is therefore updated by multiplying the new sea ice concentration, containing the increments, and the sea ice thickness. After this is done, we compute the changes in sea ice concentration and volume as a ratio (division) between the updated variables (after increments are added) and old variables (before increments are added). These changes in sea concentration and volume are used to proportionally adjust the other prognostic variables. For variables based on the volume (e.g. snow volume, ice enthalpy, melt pond volume, etc), they are multiplied by the changes in the sea ice volume, whereas for the variables based on the area (e.g. melt pond area, etc), they are multiplied by the changes in the sea ice concentration. For the melt ponds variables specifically, this proportional adjustment is done with the opposite sign, so that when DA adds ice over summer, the ponding can be reversed, avoiding the feedback issues between the DA and the melt ponds (as seen in GO6).

We have now provided more detail on some of these aspects in Section 3.1, particularly about the melt pond adjustment.

ll. 194-195, or more generally the whole discussion of 2 versus 1 covariance length scales (ll. 179-198). It would be useful to remind the reader why two covariance lengthscales was implemented, which was primarily to correct the near-surface drifts seen in the southern ocean (Figure 2e-f) and Mirouze et al. [2016].

This reminder has been added to the paper:

"The large-scale error covariances have a 400 km length-scale for temperature, unbalanced salinity and unbalanced SSH, which are demonstrated to correct near-surface drifts, particularly for salinity in the Southern Ocean (see Mirouze et al., 2016)."

• In light of only near surface differences between the 1T-1S and 1T-2S covariance lengthscale schemes, could a near surface only (damped/cutoff with depth) long length scale S covariance work even better?

There is ongoing work to address this point. We are currently testing changes to the horizontal salinity length-scales by reducing the effect of the long length-scale below the surface layers which is showing some promise. If successful, this will be included in a future update to the FOAM system but is not being implemented as part of the changes described in this paper.

• The only difference in T-profile statistics seems to be a very deep (> 1500m) increase in global RMSE of 1T-1S over 1T-2S (Figure 2a). This perhaps contradicts my statement immediately above. But more for the sake of curiosity than anything else, which region is responsible for this increase – as it is not the southern ocean.

The global RMSE increase at depths greater than 1500 m with 1T-1S configuration is coming from the Labrador Sea. When testing the length-scale setups, we found that the Labrador Sea is very sensitive to how T/S corrections are applied horizontally. Having the 1T-1S configuration gives the worst results when compared to 1T-2S or 2T-2S. Even the 2T-2S configuration produces numerical instabilities and erroneously triggers deep convection, clearly degrading the OmB statistics when compared to the run applying the Brunt-Vaisala rejection algorithm (see corrected Fig. 4d in the paper).

ll. 332-333, Section 4.2. It likely should be mentioned that the SLA degradations found in coastal regions and enclosed seas are due to model changes (Figure S7c), although degradations in the central and eastern North Atlantic, or Mediterranean outflow regions would be due to data assimilation changes (Figure S7e).

This has been added to the paper in Section 4.2:

"We also note that the SLA degradations found in some coastal regions and enclosed seas in Fig. 5 are due to model changes, whereas SLA degradations in the central and eastern North Atlantic are due to DA changes (see Fig. S7)."

• The Figures 5c/d and S7e/f should be identical, but are not? Why? In particular, the degradations in the Mediterranean Outflow region in Figure 5d seem worse than those in Figure S7f.

We believe that the reviewer missed the statement in Section S7-S9 of the supplementary material:

“It is important to highlight that the impacts of the model and DA changes on FOAM GOSI9 were evaluated in ORCA025 only, as the ORCA12 configuration is quite expensive to run for one year. However, since both configurations show very similar impacts on the observation-minus-background (OmB) statistics when comparing FOAM GOSI9 against FOAM GO6 (see Section 4 in the paper), the results below should represent a valid evaluation of the model and DA update contributions to FOAM GOSI9 improvements.”

The reason for small differences between Fig. 5c-d and Fig. S7e-f is that the breakdown of the model and DA impacts on GOSI9 was done with ORCA025 instead of ORCA12. The reason for this is that it is quite expensive to run 1-year experiments with ORCA12.

The model changes include changes to the fundamental state variables of Temperature and Salinity. Therefore, there are changes to the observation operators, and to the balance operators of NEMOVAR. Given that observation minus background (OmB) statistics are being used to show differences between the two versions:

• Can OmB statistics be fairly compared? The GOSI9 model only SST changes seem universally degraded, especially in regions of large potential T vs conservative T differences (Figure S1).

As stated in Section 4.1 of the paper, we believe the OmB statistics can be fairly compared:

“It is also worth noting that the temperature and salinity RMSD results for FOAM GO6 and GOSI9 are calculated from the EOS80 and TEOS10 variables, respectively. The magnitude of the errors is expected to be consistent whether using TEOS10 or EOS80. We investigated the impact of converting between absolute and practical salinity on the OmB values and found that it has a very small impact of the order of 0.001, which is much smaller than the salinity differences and RMSDs between FOAM GO6 and GOSI9 presented here.”

With respect to GOSI9 OmB SST statistics being universally degraded due to model changes, we respectfully disagree. The RMSE percentage changes are quite small and mostly negligible, except for the Arctic region, where the model changes do clearly degrade the SST statistics. However, as seen in Fig. S1, the difference between potential and conservative temperature in the Arctic is much smaller when compared to other regions, such as the subtropical and equatorial regions.

• Can changes to SLA OmB statistics possibly be attributed to changes in the dynamic height balance operator? Like in the coastal and enclosed sea regions?

The inclusion of the mixed layer in the dynamic height balance is not the main reason for improvements in the SLA OmB statistics. It has a large impact on the Mediterranean Outflow SLA statistics in GOSI9, but it is still a very localised improvement. The negative impacts on the SLA OmB statistics in the coastal regions and enclosed seas are mainly due to model rather than DA changes (see Fig. S7).

Figures 6&7. A spatial map of OmB RMSE for depths between 300-700m might also be useful to visualize the improvements to the system upgrades. Note: Actual depth range was chosen somewhat arbitrarily – but differences spanning 500m depth seem to be relatively global in breadth.

We do not think these additional figures are needed. We did calculate OmB profile statistics regionally, so we could consistently compare the statistics between GO6 and GOSI9 throughout the water column (up to 2000 m) and for different ocean regions. This convincingly shows the differences at depth between GO6 and GOSI9.

In light of Mignac et al. [2022] I am curious as to why / disappointed that no evaluation of sea ice thickness was performed.

We still do not assimilate sea ice thickness operationally. Additionally, although they are different sea ice models, CICE and SI^3 have largely the same sea ice physics. Therefore, small differences in the sea ice thickness are expected between GO6 and GOSI9. For this reason, we did not include any sea ice thickness comparison in the paper and focussed only on the evaluation of the sea ice concentration.

The no profile experiments answer the systems ability to operate in the pre-Argo period (1993-2005) with the presence of satellite altimeter and SST observations. However, it does not answer to the systems ability to operate in the absence of altimeter observations prior to 1993 – or a transition from no altimeter to altimeter observations in late 1992 (as for instance shown in Figure S3), particularly in the absence of correcting T/S profile measurements.

- It may be necessary to caveat statements of the usefulness of the ocean reanalysis to a particular time period (1993 onward), as “GOSI9 will lead to more potential use of Met Office ocean reanalyses in climate studies” may have different envisioned time scales depending on the user (seasonal/decadal/century scale).

Done.

“Thus, it is expected that GOSI9 results will lead to more potential for use of Met Office ocean reanalyses in climate studies, particularly for the satellite altimetry era from 1993 onwards.”

- Nevertheless, it might be worth commenting on the possible usefulness of this GOSI9 version of the Met Office FOAM system to long, century(?) scale reanalysis – or defer that discussion to a manuscript dedicated to an ocean reanalysis.

We believe that this is beyond the scope of the paper where we focus on the impact of the changes during the altimeter period.

Presentation comments

I do not find Figure 1 particularly useful. A statement that the SLA observation error in GOSI9 is larger than that used in GO6, with the difference largely attributed to the use of 4cm measurement error might suffice. The figure itself could be demoted to the supplementary material. Personally, as mentioned above, I would rather see Figure S4 promoted to the manuscript.

We think Fig. 1 is useful as it shows a clear contrast between the magnitude of the SLA observation errors in GO6 and GOSI9. Since SLA observation errors in GO6 are much smaller than in GOSI9, this also suggests that the SLA assimilation in GO6 may be overfitting the observations, exacerbating the negative impacts of the SLA assimilation on the heat content and ocean circulation when T/S profile observations are withheld. Therefore, we prefer to keep Fig. 1 in the main article.

Profile Figures. I believe it would be useful to have profile plots explicitly (in figure) labelled with a (largish) T or S (presumably in the white space normally found from mid-depth downward on the right side of most T and S profiles. This would rectify the necessity to refer to the caption all the time (although admittedly, those more familiar with the plots will immediately identify the differing shapes of T and S profiles).

Done.

Figure 2. I had difficultly, at least on the printed page, to differentiate the '2' and '1' in the labelling. You cannot improve my eyesight, but the labels could be comfortably enlarged somewhat. Improving even further beyond that might entail changing the shape of the font used 1T-2S (helvetica) may be more distinguishable than 1T-2S (computer modern).

Done.

Figure 3. One needs a Ph.D. in colour combinations to identify the regions. Caption note would seem suffice to warn the reader, but confusion will still remain. I have no suggestions, however.

This is a tricky one. We tried different things, such as only having the coloured borders with no filling, but the figure was still polluted. After a few attempts of changing this figure, we concluded that the way this figure is presented in the paper is the best way when compared to the other alternatives that we tried.

Figure 14. Vertical span of transport strength is not sufficient to adequately distinquish the differing experiment lines. In general, the figure seems oddly arranged. I would think a 3 across set of time series above a 5 across set of Hovm¨uller diagrams would work better, allowing the vertical transport strength axis to be stretched to near the size of the horzontal time axis. For time series with observed data points, an accompanying set of correlation coefficients would be useful. I am aware such an rcoefficient would only distinquish the models ability to correctly follow the observed seasonal cycle – which in turn is largely Ekman wind driven (i.e. will largely be identical across all experiments, and the observation), but this may still be useful.

We tried to rearrange Fig. 14 as suggested by the reviewer, but we still prefer how it is presented in the paper. When rearranging the figure to the format suggested by the reviewer, the AMOC stream function plots become too small, as they consist of a row of 5 plots in this format. Therefore, our preference is to keep Fig. 14 as it is.

Regarding the correlation coefficients, we do not believe this is relevant to Fig. 14. The main goal of Fig. 14 is to show the GOSI9 impacts on the transport magnitudes and how they compare to observed transports. Furthermore, it is not ideal to calculate correlation coefficients based only on 12 numbers. Refined metrics about the GOSI9 impacts on the AMOC are planned to be shown in a future publication with a long GOSI9 reanalysis.