Response to reviewer #2

We thank the reviewer for their efforts in reviewing our manuscript. These queries have helped us to revise and improve our article.

Reviewer comments are shown below in black with the author response in red.

=== General comment

This paper conducts OSSEs with a 3D-VAR-based eddy-resolving system to compare the impacts of 12 Nadir and 2 WiSA satellites on accuracy. While OSSEs are useful for evaluating various yet-to-be-constructed observation networks, this study is limited to only three experiments: assimilating standard observations, standard observations plus 12 Nadir satellites, and standard observations plus 2 WiSA satellites. To comprehensively determine the most efficient observation networks, more diverse experiments are necessary. Additionally, it would be beneficial to include information on observation coverage and funding considerations for constructing these networks.

Our aim here was to investigate the potential impact of two specific proposed observing networks to inform the planning of ESA as they explore options for the Sentinel-3 Next-Generation Topography mission. While additional experiments may have offered further insight, the computational cost of running these high-resolution systems limits the length and number of experiments. We did however include a comparison of the impacts in our high- and low-resolution systems and further experiments in the low-resolution system to further explore the impact of correlated errors.

Moreover, OSSEs produce results which are specific to the system and observation network used and can be difficult to generalise. This is made apparent by the differing results from the coordinated experiments run by the Mercator Ocean International group (reported by Benkiran et al. 2024) which we discuss in our manuscript.

As the reviewer notes in a later comment, funding considerations are another important aspect affecting the eventual choice of which observing network to implement. Given the system-specific responses, the synergy between different observation types, and external factors affecting observing platform design and deployment, we do not think that it is practically possible to optimise an observing network using OSSEs. Instead, we aimed to determine how effectively our system would be able to assimilate observations from two specific proposed networks. This allowed us to identify issues that will affect the assimilation of real wide-swath altimeter observations.

Moreover, the manuscript uses colloquial expressions, lacks fundamental details about the data assimilation systems, and does not employ statistical tests. These elements are essential for a scientific paper. Therefore, I conclude that the current paper does not meet the criteria for proceeding to the review process and expect significant revision in the next manuscript.

It is not clear to us which expressions the reviewer regards as colloquial, but we have addressed the detailed comments below on some of the phrasing used. We believe the use of the active voice rather than passive voice is a valid choice here (and in line with journal guidelines) which makes the manuscript easier to read.

We have given an overview of the data assimilation scheme used in our system in Section 2.3 with a focus on the balances and correlation length-scales used. Detail has been added on the simulation of the observation errors and how this varies between the wide-swath and nadir altimetry, and we have also included discussion on how the resulting increments differ between the experiments to understand the differing impacts. We also refer to earlier papers describing the implementation of NEMOVAR in our global forecasting system.

To demonstrate the impact of assimilating the different observation networks, we have used the standard practice for comparing OSSEs including full field differences, effect on the bias and RMSE, along with less common metrics such as the power spectral comparisons to understand the differences between the systems.

=== Specific comment

Even in the abstract, there are grammatical errors (e.g., "now able to" in L4 and "greatest" in L10) and unclear abbreviations (e.g., SWOT, SSH, RMS). Throughout the manuscript, the descriptions are written in colloquial expressions (e.g., "we see"). Therefore, it is necessary to revise the entire manuscript to ensure scientific and objective descriptions.

We do not see an issue with the phrase “now able to”. For context, we wrote that “The launch of the SWOT … mission is bringing a step change…with 2D mesoscale structures now able to be observed over the global ocean.”. We also do not see an issue with the word “greatest” in the sentence “The impact was greatest in…”. While we could have used “largest” here that might imply a wider spatial impact, rather than a difference in the magnitude of the difference.

However, we have checked the manuscript to ensure abbreviations are defined on first use and updated accordingly.

Finally, we have used phrases such as “In Figure X, we see that…” rather than using the passive voice alternatives of “In Fig X it can be seen”. We believe this is a style choice to engage a reader.

The authors use the expression "data assimilation (DA) constraints model" in this manuscript. However, DA does not make any corrections to the model source code except when it is used for parameter estimation; therefore, this expression is inappropriate.

We cannot find this specific phrase in our manuscript, though we have used phrases such as “SSH observations…play a crucial role in constraining models of the mesoscale ocean” to mean that DA constrains the model dynamics. We certainly did not mean to suggest that the DA in some way alters the model source code. We have updated the manuscript to clarify our meaning.

L32: SWOT data in 2023 became available in early 2024.

We have updated the introduction to include this point and clarify that we were referring to near-realtime observations which will be required if we are to assimilate them into realtime analysis and forecasting systems (in our case this requires the processed observations to be available at most 48 hours after the observation time).

L37: The use of "very" and similar expressions should be avoided as they lack objectivity.

Here we are describing that wide-swath altimetry observations will be “very useful”. As this is a qualitative statement, we feel the use of “very” to emphasise how useful these data are expected to be is a natural and appropriate use of language. The experiments described in the paper are an attempt to quantify how useful wide-swath altimetry can be.

This study focuses only on the two SSH observation networks (two WiSA and 12 Nadir satellites) planned by the ESA. However, the observation coverages and funding required to construct these networks are substantially different. Even if the ESA plans are currently limited to these two networks, additional sensitivity experiments are necessary to determine the most efficient observation network. Since OSSEs enable the evaluation of various unconstructed observation networks, it is essential to leverage this advantage.

The funding required to build and operate these satellites is commercially sensitive and not known to the authors. However, this was a project initiated by ESA specifically to address these two proposed scenarios which are being considered by the mission advisory groups. As mentioned above, while additional experiments may have offered further insight, the computational cost of running these high-resolution systems limits the length and number of experiments. We did however include a comparison of the impacts in our high- and low-resolution systems and further experiments in the low-resolution system to further explore the impact of correlated errors.

L46: Toy models and low-resolution models such as Lorenz-96 are used in the nature run.

It is unclear to us what the reviewer is referring to here. In OSSEs, a nature run is generally the highest resolution ocean model available and is used as a representation of the true ocean. In our experiments, the nature run is a 1/12 degree global free-running model as described in Section 2.1.

The "control run" in this manuscript is included in the OSSEs. It would be better to incorporate the control run into the OSSEs and avoid using the term "control run" throughout the manuscript.

We think our phrasing on line 52 where we referred to a control run and an OSSE run might have caused this confusion. Instead, we now refer to an OSSE framework of a control run along with additional experiments. We have retained the term “Control run” as this is a standard term to refer to the baseline experiment before the addition of more observations.

L74: Better to add 3D-VAR based before NEMOVAR.

We have updated the text to clarify that we have used a 3D-Var version of NEMOVAR.

Please specify the major differences between the NEMO models used in the nature run and the OSSEs in the 2nd paragraph of subsection 2.1.

We have updated this section to list some of the major differences between the NEMO versions.

Please add a citation for "the real-time atmospheric analysis produced at ECMWF" in L104-105.

As we can find no publication describing the specific ECMWF IFS product used, we have added a footnote with the URL linking to the ECMWF real-time data.

To confirm whether the data assimilation systems are functioning correctly, it is essential to show the prescribed observation error variance and covariance. In this manuscript, however, there are only citations of previous papers and almost no specific information. This also applies to background observation errors.

Our aim here was to estimate the impact of assimilating two specific extensions to the observing network in our operational system. To do this we constructed an OSSE to reflect that operational system and made no changes to the data assimilation scheme. In Section2.3, we give an overview of the data assimilation scheme with a focus on the balances and correlation length-scales used. Detail has been added on the simulation of the observation errors and how this varies between the wide-swath and nadir altimetry, and we have also included discussion on how the resulting increments differ between the experiments to understand the differing impacts. We also refer to earlier papers describing the implementation of NEMOVAR in our global forecasting system to avoid repeating a full system description which has already been published. We have also expanded section2.4 to include a comparison of the innovation statistics from the OSSE control and our operational system (rather than comparing operational innovation statistics at observation locations with full-field statistics from the OSSE, as we had previously). We included this comparison to demonstrate that the data assimilation system was functioning similarly in the OSSE Control and in our operational system.

Please specify “since we do not … Sentinel altimeters” in L120-121.

We have updated the text as follows to clarify our meaning. “Other satellite altimeters are also likely to be producing data at the same time as S3-NG and Sentinel-6, but since we do not know their likely characteristics we focus on Sentinel-6 in conjunction with either 2 wide-swath or 12 additional nadir altimeters.”

Since observation coverage significantly impacts the analysis accuracy, it is essential to indicate the differences in observation coverage (percentage) among the OSSEs.

Thank you for this suggestion. We have updated Section 2.2.2 to detail the number of altimeter observations assimilated in each experiment to augment the figures detailing the spatial and temporal sampling of the different observing networks. Our Control experiment

assimilated on average 188k altimeter observations per day. With the super-obbing applied to the wide-swath altimeter observations, our 2WISA experiment assimilated on average 970k altimeter observations per day (including Sentinel-6, the two wide-swath altimeters and the nadir altimeter component of each wide-swath altimeter). On the other hand, the NADIR experiment with Sentinel-6 and an additional 12 nadir altimeters assimilated on average 831k altimeter observations per day.

Please modify the description in L178-180 for readers to understand.

We have rephrased this as follows to clarify.

"The FOAM system uses a 1-day assimilation window, meaning that an analysis is produced daily using observations over a 24-hour period. The observation operator in NEMO is used to calculate a model counterpart to every observation at the nearest model timestep and interpolated to the observation location. The innovations (the difference between the observation and the model counterparts) are used by NEMOVAR together with gridded information about the model state for use in estimating the multivariate balance relationships, and information about the background and observation error covariances. The analysis increments generated by NEMOVAR (the corrections to the model state) are then read into another run of NEMO over the same day, during which a fraction of the increments are added in on each time-step using Incremental Analysis Updates (IAU; Bloom et al., 1996)."

In the third paragraph of subsection 2.4, it is unreasonable to compare the accuracy between the practical operational systems of FOAM and virtual OSSEs because these frameworks are completely different. It is unnecessary to compare these results, and it would be better to remove them.

We have improved this section by making a comparison of the innovation statistics from the OSSE control and our operational system (rather than comparing operational innovation statistics at observations locations with full-field statistics from the OSSE, as we had previously). We included this comparison because the aim of the OSSE was to emulate our real system as we are quantifying the impact in the simulated system to provide an estimate of the impact on the real system.

Please specify "incomplete" observation sampling in L214.

We have rephrased this to clarify our point that OSSEs have the advantage of knowing the true state at all model grid points and times unlike in reality where our knowledge of the true ocean state is limited.

"significant" and "significantly" can be used only if the statistical tests are conducted.

To avoid confusion we have rephrased where we previously used the term significant when referring to clear differences without specific statistical tests.

In this paper, the objective is to evaluate the impacts of 12 Nadir and 2 WiSA satellites on accuracy. However, most figures, especially those depicting spatial patterns, do not illustrate their differences, which is inconsistent with the stated objective.

We have chosen a variety of ways to illustrate and quantify the difference between our three experiments (Control, NADIR and 2WISA). This includes line plots of different metrics for the three cases and often then a direct comparison of the improvement with respect to the baseline scenario (Control). For figures depicting spatial differences, we have opted to show maps of the baseline metric (for example, the SSH RMSE from the Control) and then the change in that metric in the NADIR and 2WISA experiments. While further plots could have been included to show the difference between those differences, we did not feel this was necessary.

Please provide an explanation for why both the 12 Nadir and 2 WiSA experiments result in degraded SSH accuracy around the Antarctic region.

Thank you for highlighting this. Synthetic SSH data was not assimilated anywhere where the model sea-ice fraction was greater than 5% to emulate the operational situation. However, SSH increments due to balanced changes from temperature and salinity were spread under the ice from observations near the ice edge. While this emulates what happens in our operational, the detrimental impact on SSH under the sea-ice had not been noted in our operational system (due to a lack of observation). These experiments have highlighted that we should restrict the spreading of this information under the ice. Section 3.1 has been updated with the following text.

“Even though we have no SSH observations in sea-ice covered areas, the long background error correlation length-scale produces changes to the (highly variable) SSH under the sea-ice. While this emulates what happens in our operational system, these experiments have highlighted that we should restrict the spreading of this information under the ice.”

Please modify the descriptions in Lin 231-233.

We have updated this section (3.1) to better illustrate the effect of the different sampling of the nadir and wide-swath altimeter observations. We have included a figure showing maps of the SSH increments from each experiment on a single day and also the RMS of the SSH increments over the 21-day repeat cycle of the wide-swath altimeters. The relatively wide spacing of the altimeter swaths in the 2WISA experiment over our 1-day assimilation window produces short length-scale increments near the observation locations and longer length-scale barotropic SSH increments in the regions between altimeter swaths. In contrast, the relatively close spacing of the altimeter tracks from the 13 nadir altimeters (in the NADIR experiment) over our 1-day assimilation window produces predominantly small-scale SSH increments. The long-term effect of this is apparent in the RMS of the SSH increments over 21-days (the repeat cycle of the wide-swath altimeter observations) where larger RMS values indicate the assimilation scheme is introducing more variability in the 2WISA experiment than in the NADIR.

Adding SSH contours to Figure 4 would enhance clarity by illustrating the positions of fronts and eddies.

We chose not to do this because this plot shows the difference between the RMSE of two experiments over a month. While fronts and eddies clearly affect the structure seen here, there may be differences between the experiments from between tracks which are not coincident with these structures and any changes will also be smoothed over time.

Please specify the reasons for the degradation of temperature and salinity accuracies in the 12 Nadir and 2 WiSA experiments.

We have added the following explanation to Section 3.2. The balances in our data assimilation scheme allow altimeter observations of the SSH to introduce subsurface changes to the temperature and salinity. However, previous experiments have shown that the assimilation of in situ profiles and altimeter observations can sometimes work against one another (King et al. 2018). With such a large increase in the altimeter observations, the balanced changes applied to subsurface temperature and salinity may dominate over the changes due to the in situ observations leading to degradations in some regions over some depths, such as those seen for temperature below 1000m in the Gulf Stream region.

In Figure 5, it would be beneficial to include the temperature and salinity RMSEs in addition to the improvement ratio. To enhance clarity, consider specifying the use of different scales on the x-axis for each panel or using consistent scales across all panels.

Given the large difference in scale for the 4 plots, we have chosen to add a note in the caption to draw attention to the different scales used on the x-axes.

Since geostrophic velocities dominate most of the global ocean, it may not be necessary to present detailed validation results of surface currents in subsection 3.3. It would suffice to only describe that the results of sea surface currents are qualitatively similar to those of SSH.

We have included an analysis and discussion of the impact on surface currents as this is an important parameter of interest to many users. Quantifying the impact is therefore useful in our view. Also, we have shown that the impact, while broadly similar to SSH, is different.

No label for the color scale in Fig. 7.

Thank you for spotting this. We have added a label to the colourbar.

Please specify reasons for the differences in spatial patterns between SSH and surface currents (Figs. 3 and 7, respectively).

Our assimilation scheme uses linearised balance relationships to account for correlations between ocean variables. However the velocity balance is not applied close to the equator resulting in some of the differences seen in the spatial patterns of SSH and surface current impacts. This is now discussed in Section 3.3.

L267: Please specify reasons why the degradation signals are not distributed uniformly across the entire equatorial regions.

The differences are largest in the Amazon outflow and Somali current regions where the climatological background errors used may not properly account for the variability in these regions for the chosen period of the experiments.

In Figure 8, it is unnecessary to display both the monthly mean errors and RMSEs.

This was included to aid the interpretation of the change in both the mean flow and variability in this region discussed in Section 3.3.

The definition of the power spectral density (PSD) score appears not to be reasonable. It's unclear whether the PSD is calculated in the spatial or temporal directions, and the rationale behind calculating the ratio between the PSD of SSH error among OSSEs and that of true SSH is unclear. Since this definition is relevant to all descriptions in subsection 3.4, I will read the remaining descriptions at the next round.

Section 3.4 describes the use of two power spectra-based metrics which use the ratio of the spectral content of the error (the SSH difference between each experiment and the nature run) and the spectral content of the true signal (from the nature run) to determine a signal-to-noise ratio. The first (shown in Fig. 10) uses the 2D frequency-wavenumber power spectra to define a PSD score which distinguishes the resolved and unresolved scales in time and space. We have updated the description in Section 3.4 including a link to the open-source code used. The second metric (shown in Fig. 11) uses the wavenumber power spectrum to determine the limit of the scales constrained in each experiment. Section 3.4 has been updated to better define the PSD-based scores.

The reasons for using models with different horizontal resolutions of 0.25° and 1/12° should be specified. If the results from both resolutions are qualitatively the same, it may not be necessary to show the results from the 0.25° resolution.

We have restructured the text in Section 3.5 to clarify that the reason for using systems with different horizontal resolutions was to explore the impact on the two operational systems we run at the Met Office. After demonstrating the impacts are similar in the two systems, we used the lower resolution (and so computationally cheaper) system to run additional experiments exploring the impact of including correlated errors in the simulated wide-swath altimeter observations.

There are no universally accepted rules for using "/" to denote interchangeable expressions, as seen in "2/7% reduction in the u/v RMSE" in L340.

We have removed this and described the reductions explicitly.

The observation error variance is likely different among the three experiments (2WISA, 2WISA_CORR_TRIM, 2WISA_CORR), indicating a failure in this study to explore the impacts of observation covariance errors.

In all of the experiments we have described, the observation and background variances were not altered from those used in our operational system and Sections 2.3 and 3.5 have been updated to make this clear. While there are many ways in which our operational system could be adapted to make best use of new wide-swath altimeter observations, we felt that this work was beyond the scope of this current project. Our aim here was to investigate the impact of two specific observing network scenarios being assimilated into a system as similar as possible to our operational system (which is the method we are employing in our first attempt to assimilate real wide-swath data from SWOT).

Generally, the discussion and conclusion should be delineated separately. Moreover, a conclusion spanning over 2 pages is excessively long. Given that it does not succinctly summarize the results, I will review this section in the next revision.

We agree that this section was too long and have reorganised to separate into a discussion and conclusions. We have directly addressed the cause of the superior impact in the NADIR experiment and also the cause of the degradation in the north-east Pacific. In our conclusions we have also emphasised that this is not the best that can be achieved using these two sets of observations, but rather is a realistic estimate of their impact in our current system.