# Response to Reviewer #3

*Manuscript ID: ESSD-2025-694* *Earth System Science Data*
*Title: A global gridded dataset of significant wave height via fusion of multi-mission altimetry and numerical hindcast*
*Authors: Haoyu Jiang (Haoyujiang@szu.edu.cn), Hao Su*

---

***Review of the manuscript "A global gridded dataset of significant wave height via fusion of multi-mission altimetry and numerical hindcast " by Haoyu Jiang and Hao Su, submitted to the Earth System Science Data.***

***The manuscript presents a new global gridded SWH dataset produces by retrospective "fusion" of the WAVEWATCH III hindcast with satellite altimetry data. There a two version of the dataset – WW III fused with two altimeters (for consistency in time on the decadal scales) and WW III fused with multiple altimeters for higher precision at a given time and space.***

***The dataset aims to reproduce better accuracy for extremes in wave statistics and to be a more reliable source of wave data for climate studies.***

***The topic is an important and valuable for climate studies, however the manuscript characterized by missing of essential methodological and quantitative assessments of the dataset. I suggest that the manuscript can be published in the ESSD after the Major revision. I provide the list of my Major and Minor comments below.***

We thank the reviewer for the careful assessment of our manuscript and for recognizing the importance of the topic and the potential value of the dataset for climate studies. We appreciate the constructive comments and suggestions. We have carefully considered all the major and minor comments and provide detailed point-by-point responses below, together with corresponding revisions to the manuscript where appropriate. For a small number of points where we have a slightly different view, we also explain our reasoning in detail in the point-by-point responses below.

We also note that we were recently informed that the responses in the interactive discussion stage should be completed by 1 August (after that, we might have one month to revise the manuscript carefully). Because of the limited time available before this deadline, the present response mainly aims to clarify our position and identify the revisions that will be made. For comments from Comment 4 onward, where we agree that manuscript revisions are needed, we describe the planned changes in the point-by-

point responses, although not all of these changes have yet been fully incorporated into the manuscript. These revisions will be implemented in the formal revised manuscript submitted after the interactive discussion phase.

***MAJOR COMMENTS***

***Major comment 1***

***Authors declare that the method of fusion is an optimal interpolation, although the scheme is rather closer to a spatio-temporal interpolation model than to classical optimal interpolation.***

***The methodology lacks important statistics, including background and observation-error covariances, the error correlation function and the variance of observational errors. An extensive analysis of uncertainties should be provided. I also suggest naming the method in a more neutral way, e.g. "weighted time-space fusion for SWH".***

We thank the reviewer for this important comment. We agree that the method used in the manuscript is not a classical optimal-interpolation scheme in the strict data-assimilation sense, because it does not explicitly estimate or prescribe the full background-error covariance matrix, observation-error covariance matrix, or a formal error-correlation model from which statistically optimal weights are derived. The original wording "space-time optimal interpolation" may therefore give an impression of a more formal OI framework than what is actually implemented. To avoid this ambiguity, we have revised the terminology and now refer to the method as a "space-time weighted fusion" method, while explaining that it is inspired by the structure of optimal interpolation but implemented as a pragmatic weighted-increment reconstruction of SWH.

The purpose of the method is to reconstruct a gridded historical SWH field by applying observation-derived SWH increments to a dynamically complete WW3 background field. Specifically, the fused field is obtained by adding to the background SWH a weighted average of altimeter-minus-background SWH differences. The weights are not derived from a full covariance model, but are prescribed according to physically motivated measures of observational relevance, including spatial separation, temporal separation, modelled-SWH similarity, and the correlation between the background SWH time series at the target and observation locations.

This formulation is based on the assumption that, after cross-mission calibration and denoising, along-track altimeter SWH observations provide substantially more accurate

local constraints than the non-assimilative WW3 background. This assumption is supported by previous altimeter validation studies and by the validation results presented in the manuscript. The objective is therefore not to produce a minimum-variance analysis under a fully specified statistical covariance model, but to construct an observation-constrained reconstruction that retains the spatial and temporal continuity of the numerical background while making maximum use of the more accurate altimeter SWH observations.

We agree that a fully covariance-based OI formulation would in principle require estimates of background-error covariance, observation-error covariance, and representativeness error. However, these quantities are difficult to estimate robustly at the global scale for a multi-decadal retrospective reconstruction. Background errors depend on storm regime, atmospheric forcing errors, swell propagation, source-term parameterizations, sea-ice conditions, and regional wave climate. Observation errors also vary among missions, sea states, retrackers, coastal conditions, and sampling geometry, and include representativeness differences between point/track observations and grid-cell-scale SWH. Introducing poorly constrained covariance models could therefore give a misleading impression of formal uncertainty quantification.

Instead, we evaluate the uncertainty and robustness of the product empirically. The manuscript includes validation against independent buoy observations, leave-one-satellite-out validation against withheld altimeter missions, a separate assessment of the one-dimensional bias adjustment, sensitivity tests for the fusion parameters, and a background-sensitivity experiment using two substantially different WW3 hindcasts. These assessments quantify the actual performance of the final product and show that the improvement cannot be explained by a simple mean-bias correction alone.

We also note that the present product corresponds to a deliberately observation-trusting reconstruction. In applications where a more conservative use of observational increments is desired, the fused increment can be downweighted by blending the fused field with the original WW3 background. Such a procedure would not replace a full covariance-based reanalysis, but it provides a simple way to explore sensitivity to the relative strength of the observational constraint.

In response to this comment, we have revised the terminology used to describe the method and have clarified in the Methods section that the approach is a space–time weighted fusion method rather than a classical covariance-based optimal-interpolation analysis. We have also added text explaining the implicit assumption of stronger trust in calibrated altimeter SWH observations than in the non-assimilative WW3

background, and we clarify that the product uncertainty is assessed empirically through independent validations and sensitivity experiments rather than through a formal analytical error covariance model. The manuscript has been revised accordingly:

*"Prior to data fusion, a simple one-dimensional look-up table correction is applied to remove the systematic bias between SWHs from the WW3-ST6 hindcast data and the along-track altimeter observations. The data-fusion method employed in this dataset is a space-time weighted fusion approach inspired by optimal interpolation and objective analysis, originally developed for meteorological and oceanographic data analysis (e.g., Bretherton et al., 1976). In this framework, sparse observations are used to correct a background field through distance- or covariance-dependent weights. The present study modifies this framework for SWH reconstruction to give strong local influence to calibrated altimeter SWH observations within the fused product with the corresponding formula expressed as follows:*

EQUATIONS 1-4 & THEIR EXPLANATION

*This method implicitly gives strong confidence to calibrated altimeter SWH observations relative to the non-assimilative WW3 background. The uncertainty of the resulting product is therefore assessed empirically through independent buoy validation, leave-one-satellite-out altimeter validation, parameter-sensitivity tests, and background-sensitivity experiments, rather than through a formal analytical error-covariance model."*

*"Because the present method does not prescribe a formal analytical error-covariance model, the uncertainty and robustness of the fused product are assessed empirically through independent validation and sensitivity experiments. To assess the impact of background-field selection,…"*

***Major comment 2***

***I suggest to provide more technical details for the look-up table that was used for the correction of both WW III hindcast and altimeter data. How was the look-up table built, e.g., was it built globally or regionally, and what time period was used? The Jason-2 independent data for comparison excluded from the lookup table?***

***From lines 166–175 it is not clear how the look-up table is related to the exclusion of observations according to the specified thresholds. Is the observation filtering performed using the look-up table, or is it a separate processing stage? This procedure should be described more clearly and illustrated with a plot or a schematic diagram.***

We thank the reviewer for pointing out that the construction and role of the look-up table were not described sufficiently clearly. We agree that this step should be explained in more detail.

The look-up-table correction used in this study is a simple one-dimensional nonlinear bias adjustment, analogous to a linear calibration but without assuming a constant slope and intercept. It is applied only to the WW3 SWH background field before the space-time fusion. The CCI Sea State altimeter observations themselves are not corrected by this look-up table because they have already undergone cross-mission calibration and denoising in the CCI processing chain.

The look-up table was constructed globally from collocated pairs of WW3-ST6 SWH and CCI Sea State along-track SWH. The samples were grouped according to the WW3 SWH value, and the mean altimeter-minus-WW3 difference was calculated for each 0.1-m SWH bin. The resulting correction was then applied as a function of the modelled SWH. For SWH values above 12 m, where the number of collocated samples is much smaller, a simple linear-regression correction was used instead of fitting individual bins, in order to avoid overfitting in the sparsely sampled high-wave range. The correction is one-dimensional and global and does not vary by region, season, or satellite mission.

We also clarify that the look-up-table correction is independent of the subsequent observation-screening procedure. The look-up table is a pre-fusion bias-adjustment step. In contrast, the exclusion of observations according to spatial separation, temporal separation, modelled-SWH difference, and background-correlation thresholds is performed during the space–time weighted fusion stage. Therefore, the look-up table is not used to decide whether an observation is included or excluded from the fusion.

Regarding the reviewer's question about Jason-2, in the original implementation Jason-2 observations were included in the construction of the global look-up table, but were excluded from the subsequent space–time fusion when Jason-2 was used as the withheld validation reference. Thus, the Jason-2 comparison is independent with respect to the fusion step, although not completely independent with respect to this global pre-fusion bias adjustment. However, this inclusion has only a very limited effect on the validation results. The look-up table is a low-degree, globally applied SWH-dependent bias correction constructed from the jointly calibrated multi-mission CCI dataset, and it is not a satellite-specific or locally fitted correction. Moreover, the manuscript evaluates the look-up-table correction separately and shows that it reduces the RMSE by only approximately 0.01 m and increases the correlation coefficient by

only 0.001. The substantial improvement obtained after the full fusion therefore cannot be attributed to the look-up-table step or to the inclusion of Jason-2 in that global calibration. look-up table more explicitly. We have also clarified that the look-up-table correction is separate from the observation-screening procedure used in the space-time weighted fusion:

*"Prior to the space-time weighted fusion, a simple global one-dimensional look-up-table correction is applied to reduce the systematic SWH-dependent bias of the WW3-ST6 background relative to the calibrated CCI Sea State altimeter observations. This correction is analogous to a nonlinear calibration. Collocated WW3-ST6 and CCI Sea State SWH pairs are grouped according to the WW3-ST6 SWH value, and the mean altimeter-minus-WW3 difference is calculated within each 0.1-m SWH bin. The resulting bin-mean correction is then applied to the WW3-ST6 SWH field and to the model-equivalent SWH values collocated with each altimeter record. For SWH values above 12 m, where collocated samples are relatively sparse, a simple linear-regression correction is used instead of fitting individual bins, in order to avoid overfitting in the high-wave range. All subsequent data-fusion procedures are therefore based on the look-up-table-corrected hindcast background rather than on the original hindcast."*

***Major comment 3***

***The statement that Jason-2 observations were withheld from the 2020 CCI Sea State dataset appears inconsistent with the mission selection reported elsewhere in the manuscript: Table 1 indicates that Jason-3, rather than Jason-2, was used during this period. In addition, the text below Eqs. (5)–(7) refers to a fused dataset "without Jason-1" validated against Jason-1, which is inconsistent with the preceding description and appears to be another mission-name error. These inconsistencies should be carefully checked and corrected.***

***The tuning needs a quantitative description, including tested ranges and increments, whether parameters were varied individually or jointly, and the objective criterion used to combine bias, RMSE, and CC. The statement that all parameters were varied from −50% to +100% cannot be interpreted literally for the CC threshold of 0.7, since a 100% increase would produce 1.4, outside the range of a correlation coefficient.***

We thank the reviewer for carefully identifying these inconsistencies. The reviewer is correct that there were mission-name and year-description errors in the original manuscript. The tuning and leave-one-satellite-out validation experiments described in this section were conducted using the 2011 CCI Sea State data, not the 2020 data. In this experiment, Jason-2 was withheld from the space–time fusion and used as the

reference dataset for evaluating the fused result, as shown in Fig. 3c for the final result. The same leave-one-satellite-out strategy was also applied to the other available missions, and the corresponding results are summarized in Tables 2–4.

The statement below Eqs. 5-7 referring to the fused dataset "without Jason-1" validated against Jason-1 was also a typographical error. It has been corrected to refer to the corresponding withheld satellite used in each validation experiment. We have carefully checked and corrected these mission-name inconsistencies throughout the revised manuscript.

Regarding the tuning procedure, we agree that the original description was too brief. The parameter tuning was performed empirically through a series of semi-monthly (to exclude unreasonable values) and annual test experiments. The spatial scale $S_1$ was fixed at 50 km and the remaining parameters were tested over the following candidate values: $C = 2,4,8,16$; $T_1 = 30,90,180$ min; $Q = 2000,4000,8000,16000$; $S_{\mathrm{thr}} = 1000,2000,4000$km; $T_{\mathrm{thr}} = 6,12,24,48$h; and $r_{\mathrm{thr}} = 0.6,0.7,0.8,0.9$. We regard the selected parameter values as near-optimal for two reasons. First, within the tested parameter range, this parameter set yielded the smallest RMSE while avoiding visually evident satellite-track artefacts in the fused fields. Second, the resulting RMSE was already close to the level typically attainable in satellite-to-satellite comparisons. This indicates that the fusion procedure was approaching the practical accuracy limit imposed by the available altimeter observations and their sampling characteristics, with the RMSE reaching approximately 0.25 m.

At the same time, after reconsidering this point, we prefer not to include an extensive table or figure showing all tuning experiments in the manuscript. The parameter tuning in this study was not intended as a formal optimization problem or as a major methodological contribution. Rather, it was a pragmatic sensitivity check used to select a reasonable set of parameters for the space-time weighted fusion. The tested parameter combinations mainly served to avoid obviously undesirable behavior, such as excessive along-track imprints, while maintaining low RMSE, small bias, and high correlation against withheld altimeter observations.

The sensitivity tests also showed that the fused results were generally insensitive to moderate changes in the parameters near the selected values. This limited sensitivity is physically understandable. In most cases, the target grid point is constrained simultaneously in several dimensions, including spatial separation, temporal separation, modelled-SWH similarity, background-field correlation, threshold-based observation selection, and the base-distance term. Therefore, the fusion is generally an interpolation

of altimeter-minus-background increments among multiple dynamically related observations, rather than an extrapolation controlled by a single tuning parameter. One notable exception was $C = 2$, for which the resulting fields showed overly strong along-track imprints of the altimeter observations. Apart from such clearly unreasonable cases, moderate changes in individual parameters had only a limited impact on the final fused SWH field.

To balance reproducibility and conciseness, we have revised the Methods section to state the final parameter values, the validation year, the withheld-satellite strategy, and the empirical selection criterion. We have also removed the imprecise statement that all parameters were varied from $-50\%$ to $+100\%$, because this wording is not meaningful for the correlation threshold. We believe that this revised description provides the necessary information for understanding how the parameters were selected, while avoiding an overly detailed presentation of implementation-level tuning experiments:

*"Here, C, $S_1$, $T_1$, Q, $S_{thr}$, $T_{thr}$,* and $r_{thr}$ *are set to 8, 50 km, 90 min, 8000, 2000 km, 24 h, and 0.7, respectively, based on a straightforward but trivial tuning procedure which was conducted using the 2011 CCI-Sea State dataset. Specifically, Jason-2 data were withheld, and the remaining three altimeter missions (CryoSat-2, ENVISAT, and Jason-1) were fused with the WW3-ST6 hindcast data. The fused results were then validated against the independent Jason-2 observations, and the combination of tuning parameters (C, $S_1$, $T_1$, Q, $S_{thr}$, $T_{thr}$, $r_{thr}$) that yielded relatively small overall validation errors was selected. The error metrics considered in this evaluation include bias, RMSE, and correlation coefficient (CC):*

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\right) \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\right)^2} \tag{6}$$

$$CC = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)\left(x_i - \bar{x}\right) \Big/ \left[\sqrt{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right] \tag{7}$$

*where x and y denote the SWH from the dataset to be evaluated and the reference data (in this case, fused dataset without Jason-2 and Jason-2 from CCI-Sea State) , respectively; n is the sample size, and the bars over them denote their mean values. It is worth noting that the fused results are not sensitive to these tuning parameters near their optimal values. Moderate changes in these tuning parameters only weakly affect the fused results. This is because most target grid points are simultaneously constrained by spatial separation, temporal separation, modelled-SWH similarity, background-field correlation, threshold-based observation selection, and the base-distance term. The*

*fusion therefore mainly represents an interpolation of altimeter-minus-background increments among dynamically related observations, rather than an extrapolation controlled by a single parameter."*

We believe that this revised description provides the necessary information for understanding how the parameters were selected, while avoiding an overly detailed presentation of implementation-level tuning experiments. To further ensure transparency, we can also provide the complete set of tuning experiments and their corresponding validation statistics as supplementary material, should the reviewer or editor consider this necessary.

***Major comment 4***
***The independence of the Jason-2 data for validation is questionable, as the CCI Sea State data are intercalibrated across missions. While the leave-one-out validation is a common and appropriate approach, Jason-2 is not fully statistically independent. This needs to be explicitly clarified.***

***Also, was Jason-2 excluded from all processing stages? Look-up table correction, parameter tuning?***

We thank the reviewer for raising this important point. We agree that Jason-2 should not be described as fully statistically independent in the strictest sense, because all altimeter missions in the CCI Sea State dataset have undergone cross-mission calibration and denoising within a common processing framework. We will clarify this point in the revised manuscript.

At the same time, cross-mission calibration does not make the Jason-2 observations unusable for validation. Its purpose is to reduce systematic inter-mission biases (via $y=kx+b$) and improve the consistency of the altimeter record, rather than to reconstruct Jason-2 measurements from the other satellites. The Jason-2 along-track observations retain their own sampling pattern, measurement noise, sea-state dependence, and space–time coverage. Therefore, withholding Jason-2 from the space–time fusion still provides a meaningful test of whether the fused field, constrained by the remaining satellites and the WW3 background, can reproduce independent along-track SWH observations not directly assimilated in the fusion step.

We also clarify the role of Jason-2 in the different processing stages. When Jason-2 is used as the reference in the leave-one-satellite-out validation, it is excluded from the space–time weighted fusion. Jason-2 was also included in the construction of the simple

global one-dimensional look-up-table correction. As clarified in response to Major Comment 2, this look-up table is a low-degree, globally applied SWH-dependent bias correction used only to reduce the systematic difference between the WW3 background and the calibrated CCI altimeter record. It is not satellite-specific and is not used to introduce local Jason-2 information into the fused field.

More generally, we note that complete statistical independence is difficult to achieve in satellite altimetry validation. If any dataset that has undergone cross-calibration against other observing systems were considered unsuitable for validation, then truly independent reference data would be difficult to obtain. Even in situ buoy records, which are often treated as independent references, have been widely used in the calibration and validation of satellite altimeter SWH products. Therefore, the relevant issue is not whether the validation dataset is absolutely independent of every upstream calibration activity, but whether it is independent of the specific processing step being evaluated. In the present case, Jason-2 is not independent of the common CCI calibration framework, and we now explicitly state this limitation. However, it is independent of the space–time fusion step being evaluated, because its along-track observations were not used as local constraints in that fusion experiment.

***Major comment 5***
***The justification for the two-satellite version of the dataset is the temporal homogeneity of the data for climate studies, which I fully support. However, the constant number of altimeters does not guarantee the homogeneity of the time series. It should be demonstrated that the replacements of TOPEX/Poseidon/Jason and ERS/Envisat/CryoSat do not introduce artificial trends or inconsistencies in the fused dataset. A dedicated analysis of possible discontinuities at satellite mission transitions is still required. A simple analysis may be added in a separate figure:***
***- the time series of global and regional mean SWH and differences A−M***
***- trends in the two-satellite dataset***
***-a global map of SWH trends.***

***The analysis should be provided both globally and regionally, with a focus on the polar regions.***

***Although the authors state that the initial CCI Sea State v4 data have undergone extensive inter-mission calibration for consistency, this does not directly imply the temporal homogeneity of the resulting two-satellite fused product. I would suggest providing basic time series and global and regional plots for the two-satellite dataset for the whole period, 1992–2020, in order to identify possible discontinuities at satellite mission transitions.***

We thank the reviewer for this important suggestion. We completely agree that maintaining a constant number of altimeters does not by itself guarantee temporal homogeneity of the resulting two-satellite product. The replacement of satellite missions may still introduce artificial discontinuities through differences in orbit, sampling geometry, instrument characteristics, or residual inter-mission calibration uncertainties. This issue deserves explicit assessment, particularly because the two-satellite product is intended for wave-climate applications.

In response to this comment, we will add a dedicated analysis of possible discontinuities associated with satellite mission transitions in the two-satellite product. Specifically, we will examine the main transition times in the selected satellite pairs, such as the TOPEX/Poseidon–Jason-series transitions and the ERS–Envisat–CryoSat-related transitions. For each transition, we will analyse time series of global and regional mean SWH, as well as the corresponding fused-minus-background difference before and after the transition. The analysis will be performed not only globally, but also for selected regional domains, with particular attention to high-latitude and polar regions where sampling geometry, sea-ice effects, and mission coverage may introduce larger uncertainties.

The purpose of this additional analysis is to determine whether mission replacements introduce detectable step-like changes or abrupt shifts in the fused product that cannot be explained by the WW3 background variability or by natural wave-climate variability. We will therefore focus on transition-centred diagnostics, such as before–after differences in regional mean SWH and $A - M$, rather than relying only on long-term trends. This approach is more directly targeted at identifying artificial discontinuities caused by satellite replacement.

We will revise the manuscript accordingly by adding a new figure and accompanying discussion. This will clarify the temporal behaviour of the two-satellite product and provide a direct check for possible discontinuities associated with satellite replacements.

***Major comment 6***

***I don't quite get why all comparisons are performed for a single year, 2011? Authors state that 2011 was chosen to reduce the volume of the data processing. At the same time, producing the dataset for 1992 – 2020 period already took the data processing time and effort. Therefore, the argument of reducing the data-processing volume is not convincing enough to justify validation based on only one year.***

***The authors make conclusions about the accuracy, performance, and advantages of the resulting dataset, while almost all scatter plots, maps, and statistical comparisons are presented only for 2011. It is not clear whether this particular year is representative of the complete period. The wave climate may vary significantly between different years. For a dataset covering almost three decades and intended for climate applications, validation based on a single year seems insufficient.***

We thank the reviewer for this important comment. We agree that the original explanation that 2011 was selected simply "to reduce the volume of data processing" was insufficient and could give the impression that the validation was based on an arbitrary single year. We will revise the text to clarify the rationale for selecting 2011 and to better distinguish between validation of the fusion method and validation of the long-term dataset.

The validation strategy is based on the assumption that both components entering the fusion framework are temporally consistent over the study period. The WW3-ST6 hindcast provides a dynamically consistent background generated using a fixed model configuration, while the CCI Sea State observations provide a cross-calibrated and denoised multi-mission altimeter record. Under this assumption, the one-year intensive leave-one-satellite-out experiment is intended to evaluate the behavior and achievable accuracy of the fusion method itself, rather than to re-estimate the method performance independently for every year. In other words, if the background and observational inputs remain broadly consistent, the main error characteristics of the fusion scheme should not fundamentally change from year to year, although the specific error statistics may slightly vary with wave climate, satellite sampling, and regional sea-state conditions. Nevertheless, we agree that this assumption should be checked for the long-term two-satellite product, especially at satellite mission transitions, as replied before.

The year 2011 was selected primarily because four altimeter missions were available simultaneously, allowing a relatively comprehensive leave-one/two-satellite-out combination validation. This provides a useful test of the stability of the fusion method under different combinations of assimilated and withheld satellites. The results shown in Fig. 3 and Tables 2–4 were therefore designed mainly to evaluate the methodological performance of the fusion scheme, rather than to claim that a single year fully represents the entire 1992–2020 wave climate.

We also note that generating the final dataset once is not the same as performing repeated leave-one/two-satellite-out validation experiments. Each validation experiment requires a new fusion run with a different satellite configuration, followed

by along-track collocation with the withheld satellite and comparison among different satellite combinations. Therefore, a complete repetition of all combination-based validation experiments for every year would involve substantially more processing than the one-time production of the dataset itself.

In addition, the manuscript does not rely only on the 2011 satellite-based validation. The buoy validation was performed using 50 offshore NDBC buoys over the period 2010–2016, which provides a multi-year independent in situ assessment of the fused product. This period was selected not only to provide sufficient validation samples, but also to maintain a relatively consistent buoy reference. Extending the validation over a much longer period may introduce additional inhomogeneity because the NDBC buoy network has experienced changes in instrumentation, processing, and/or reported measurement precision over its long record. Such changes could complicate the interpretation of small differences in validation statistics. The satellite-based leave-one-out experiment and the buoy validation therefore serve different purposes: the former is a controlled test of the fusion algorithm under multiple satellite-combination scenarios, while the latter evaluates the dataset against independent in situ observations over a longer but relatively consistent period.

We agree that wave climate varies from year to year. However, the effect of interannual variability on the fusion error metrics is not straightforward to isolate, because changes in validation statistics may arise not only from wave-climate variability, but also from changes in the satellite constellation, sampling geometry, number of available missions, and regional observation density. For some years, especially when only two relevant altimeters are available, a leave-one-satellite-out experiment would also become less meaningful.

Nevertheless, we agree with the reviewer that the manuscript should not overstate conclusions drawn from the 2011 satellite-validation experiment alone. We will therefore revise the wording to clarify that the 2011 experiment is an intensive methodological validation case enabled by the availability of multiple simultaneous altimeters. We will also add the mission-transition consistency analysis described in response to the previous comment.

***Major comment 7***
***The manuscript does not provide a single table with the main characteristics of the dataset. The following information should be clearly presented:***
***- the exact period of each product version;***
***- the spatial resolution;***

***- the temporal step;***
***- the rules used for filling missing values;***
***- the version of WW3-ST6 used;***
***- the total dataset volume;***
***- the available variables.***

***The manuscript states at the same time that the product keeps the original NWM resolution of 0.25° and that its native resolution is 0.5°. The CCI Sea State v4 dataset is described as starting in August 1991, while the two-sat product in the table starts in October 1992. The terms "multi-sat" and "all-sat" are also used in different sections. These inconsistencies should be corrected.***

We thank the reviewer for this helpful suggestion. We agree that the main characteristics of the released dataset should be summarized more clearly in the manuscript. In the current version, some of this information is already included as metadata in the NetCDF files, such as the source WW3-ST6 hindcast file, the fused satellite missions, the start and end time of each file, satellite identifiers, and the relative contribution of each satellite mission to the fusion. This organization is useful for data users working directly with the files. However, we agree that relying on NetCDF metadata alone is not sufficient for a data-description paper, and that the manuscript itself should provide a concise and complete overview of the released products.

We will add a table in the revised manuscript summarizing the key properties of each product version. This table will include the exact temporal coverage, spatial resolution, temporal resolution, available variables, missing-value convention, background WW3-ST6 information, and approximate total data volume. We agree that such a table will make the dataset easier to understand and use.

Regarding the apparent inconsistency in spatial resolution, we clarify that the original WW3-ST6 hindcast used as the background field is available at 0.25° spatial resolution, whereas the released fused dataset is provided on a regular 0.5° grid. The statement that the fused dataset "preserves the original resolution of the NWM" was imprecise and may have caused confusion. Our intended meaning was that the fused product preserves a regular, spatially complete model-based grid structure, rather than being a sparse or track-dependent gridded altimeter product. We will revise this wording to state explicitly that the native resolution of the released fused product is 0.5°.

We will also clarify the difference between the temporal coverage of the input CCI Sea State v4 dataset and that of the released fused products. Although the CCI Sea State v4

record starts in August 1991, the fused product starts later because the construction of this version requires at suitable pair of altimeter missions satisfying the prescribed sampling configuration. Before October 1992, there is only one satellite (ERS-1) in orbit. This distinction will be stated explicitly in the new version.

We also realized the terminology problem after submission. The previously use of "all-sat" have been corrected throughout the manuscript to "multi-sat".

***Major comment 8***

***A more detailed description of the observation filtering near sea ice is required. These are exactly the regions where the largest errors remain in Figure 4.***

We thank the reviewer for pointing this out. We agree that the treatment of observations near sea ice should be described more explicitly.

In the present version of the dataset, no dedicated sea-ice-specific correction or dynamic ice-edge filtering scheme was applied beyond the quality control and availability constraints inherited from the input datasets and the general observation-screening criteria used in the fusion. Therefore, grid points close to sea ice should be interpreted with caution. In these regions, the WW3 background field is generally less reliable because wave evolution can be strongly affected by sea ice, marginal ice-zone processes, and possible inaccuracies in the ice mask. At the same time, valid altimeter observations are sparser and may be located farther away from the target grid point, so the fusion can become less directly constrained by local observations and more dependent on extrapolation from nearby open-water conditions.

This explains why relatively large residual errors remain near sea-ice margins even after fusion. The issue does not indicate a failure of the fusion method in the open ocean, but rather reflects the reduced reliability of both the numerical background and the observational constraints in regions that are strongly affected by sea ice.

In response to this comment, we will revise the Methods section to state more clearly how sea-ice-affected regions are treated in the current processing. We will also add a caution in the Data Description section noting that the fused SWH values near sea-ice margins should be used with particular care, and that the product is primarily intended for open-ocean applications where both the WW3 background and the altimeter constraints are more reliable.

***Minor comments***
***Section 2.1 proceeds from Sect. 2.1.2 directly to Sect. 2.1.4.***

***Consistent name for the multi-satellite product throughout the manuscript. Both "multi-sat" and "all-sat" are currently used.***

***"the parameter set yielding relatively smallest errors" is grammatically and methodologically unclear. Please specify the exact optimization criterion.***

***Equations 5-7: please define explicitly which dataset is represented by x and which by y.***

***Please describe the temporal collocation between the 20-minute NDBC buoy observations and the gridded data.***

***A compact table summarizing the final datasets, including temporal coverage, spatial resolution, temporal resolution, variables and approximate data volume is needed.***

We thank the reviewer for these careful minor comments above. We agree with all of them and will revise the manuscript accordingly. These corrections will be incorporated into the revised manuscript.

***The number of observations used in each validation experiment should be reported.***

We thank the reviewer for this suggestion. We clarify that the number of valid matchups has already been shown in the main scatter plots. However, we understand that the reviewer might refer to the satellite-combination validation experiments summarized in Tables 2–4, for which the sample sizes corresponding to each fused-satellite pair and withheld reference mission were not explicitly tabulated. We agree that this information should be provided for completeness and reproducibility. In the revised manuscript, we will add the number of valid collocated observations used in each satellite-combination experiment by adding a table.

***The improvement of approximately 0.005 m is very small. Please state whether this difference is statistically significant.***

We thank the reviewer for this comment. We agree that the RMSE reduction of approximately 0.005 m is very small in absolute terms and should not be over-interpreted as a practically large improvement in the global error metric. Given the large

number of valid collocations (more than 8,000,000), this difference is statistically significant.

Nevertheless, the SWH-based distance term still has a useful physical role. Its purpose is not only to reduce the domain-mean RMSE, but also to improve the behavior of the fusion in cases where the modelled wave-height field and the altimeter observations are spatially or temporally displaced. Such displacement may occur when a modelled storm or swell system has an arrival-time or position error. In these situations, the SWH-based distance helps assign relatively larger weights to observations that are more likely to belong to the same wave system, and lower weights to observations from dynamically different sea states, even if they are geographically or temporally close.

We will revise the manuscript to clarify that the overall RMSE reduction associated with this term is small, while its main motivation is to improve the physical consistency of local correction increments under model–observation displacement conditions.

***Lack of jointly calibration -> Lack of joint calibration***

***It provide hourly data many wave parameters -> It provides hourly data for many wave parameters***

***grided -> gridded, meterological -> meteorological, Ribl & Young -> Ribal and Young***

***The statement that the bias is "statistically negligible at the climate scale" should be supported quantitatively or rephrased as "small relative to the mean SWH"***

All corrected (the above four comments).

***The Code and Data Availability section should distinguish clearly between input datasets and the final fused products.***

We thank the reviewer for this suggestion. In the manuscript, the final fused products are described in detail in the dedicated "Data Description and Availability" section, where the two released product versions, the stored variables, the NetCDF organization, and the associated metadata are introduced. The separate "Code and Data Availability" section was intended to provide source information for all datasets used in the study, including the input altimeter data, WW3 hindcasts, comparison datasets, buoy observations, and the final fused dataset.

In the revised manuscript, we will reorganize the Code and Data Availability section to clearly separate: (1) input datasets used for fusion, including the CCI Sea State altimeter observations and WW3-ST6 hindcast; (2) reference datasets used for validation and comparison, including NDBC buoys, CMEMS gridded SWH, and WW3-ST4; and (3) the final fused products released in this study. We will also add cross-references to the Data Description and Availability section, where the final product characteristics are described in more detail.

***Figure 4 caption – which time period?***

We thank the reviewer for pointing this out. Figure 4 is based on the same validation experiment as Figure 3, namely the comparison against withheld Jason-2 observations in 2011. We will revise the Figure 4 caption to state the time period explicitly.

***Why Figure 7 demonstrates results for 2020?***

We thank the reviewer for pointing this out. Figure 7 is based on the results of Wang and Jiang (2024), for which 2020 was used as an independent test year. The purpose of including this figure is to illustrate a downstream application of the fused dataset in AI wave modelling, rather than to provide an additional validation year for the fusion method itself. Therefore, the use of 2020 follows the experimental design of the original study. We agree that this was not sufficiently clear in the manuscript, and we will revise the text and caption of Figure 7 to state explicitly that the results are shown for the independent 2020 test year used in Wang and Jiang (2024).

Again, we sincerely thank the reviewer's detailed assessment of our manuscript. The comments have helped us identify several places where the methodology, validation design, dataset description, and terminology should be clarified. We are grateful for these constructive suggestions and will implement the corresponding revisions carefully in the revised manuscript.