Response to Reviewers

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

Dear Editors,

Thank you for consideration of our manuscript for publication in *Earth System Science Data*. We appreciate you and the reviewers for your earnest work. The comments from the reviewers are helpful, and the paper has been revised carefully according to some of these comments.

Please find attached the point-by-point response to the reviewers' comments, as well as a tracked-changes version of the manuscript. For comments where we may have slightly disagreed, we have also provided our explanations in the response.

We hope that this version of the manuscript meets the high standards of *Earth System Science Data* and is now acceptable for publication. We look forward to your further assessment and are happy to address any additional questions.

Thank you again for your time and support.

Best regards,
Haoyu

Response to Reviewer #1

Review of "A global gridded dataset of significant wave height via fusion of multi-mission altimetry and numerical hindcast"
a paper by Jiang and Su

Synopsis
The authors evaluate the "data fusion" method for blending SWH fields from wave model(s) and altimetry. The method and dataset (or a subset of the dataset, at least) has been previously published. I interpret the methods as a kind of enhanced optimal interpolation. The authors find high accuracy with their product. They offer two forms of their product. One prioritizes temporal stability, which is useful for climate studies. The other prioritizes general accuracy (lower RMSE). The authors encourage others to download and use their dataset.

Recommendation: I think this should be published, after some modest revision.

We sincerely thank the reviewer for the positive assessment of our manuscript and for the constructive and detailed comments. We are pleased that the reviewer considers the dataset suitable for publication after modest revision. The manuscript has been revised according to the comments of the reviewer. The authors hope that the revised version of the manuscript is acceptable to the reviewer. Detailed responses are provided below.

General remark

What is the original reference for the data fusion method? I'm not familiar with the history of the method, but lines 430-432 indicate that there was a previous publication, Wang and Jiang 2024, in which a subset of this same data-fusion dataset was used. I would expect this to come up in Section 2.2, but I didn't find it there.

We thank the reviewer for pointing this out. We agree that the original manuscript did not clearly distinguish between the methodological background of the fusion approach and the previous use of a subset of the fused dataset in Wang and Jiang (2024).

The fusion method used in this study is not intended to be presented as a completely new method. It is based on the general framework of objective analysis / optimal interpolation, for which one classical reference is Bretherton et al. (1976). In the present study, this framework is modified for gridded SWH reconstruction by incorporating several wave-specific factors into the distance/weighting definition, including temporal separation, spatial separation, modelled-SWH difference, correlation-based screening, and a base-distance term to avoid unrealistic along-track striping caused by overly large weights near altimeter tracks.

We have revised Section 2.2 to explicitly cite Bretherton et al. (1976) and to clarify that the present method is a modified space-time objective-analysis approach adapted for SWH fusion. We have also clarified the role of Wang and Jiang (2024): that study used only a five-year subset of the fused SWH product for AI wave-model training and fine-tuning, but it did not provide a dedicated description, validation, or public release of the full long-term gridded dataset. The present manuscript is therefore the dataset paper, providing the full methodology, systematic validation, two product configurations, and data-access information.

Here are the revised texts:

"*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.*

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"

"*It is noted that Wang and Jiang (2024) focused on AI model development and used the fused SWH fields only as training data. It did not provide the full long-term dataset, the two product configurations, the systematic validation, or the public data documentation presented here.*"

Added Reference:

Bretherton, F. P., Davis, R. E., and Fandry, C. B.: A technique for objective analysis and design of oceanographic experiments applied to MODE-73, Deep-Sea Research, 23, 559–582, 1976.

Comments

Line 14: is there a distinction between "fusion" and "post-fusion"?

We agree that the terminology was potentially confusing. By "post-fusion" we intended to emphasize that the correction is performed offline after the numerical hindcast has been produced, rather than through online data assimilation inside the wave model. To avoid ambiguity, we have replaced "post-fusion" with "offline fusion" throughout the manuscript.

Line 86-92: It is not clear from this text if all three limitations are addressed.

We thank the reviewer for this helpful comment. We agree that the original text did not explicitly state how the proposed dataset addresses the three limitations discussed in the preceding paragraph. We have revised the Introduction to clarify that the first two limitations are directly addressed by using jointly calibrated CCI Sea State altimeter observations and by applying an offline fusion framework that can use both past and future observations. The third limitation is avoided, rather than directly solved, because our product reconstructs SWH only and therefore does not require modification of the full directional wave spectrum. The revised text reads:

"*The proposed fusion framework addresses above limitations in different ways. First, it uses the jointly calibrated CCI Sea State altimeter record as the observational reference, thereby avoiding calibration against the model background. Second, because the fusion is performed offline after the hindcast has been generated, both past and future altimeter observations can be used to constrain the SWH field at a given analysis time. Third, this retrospective constraint also reduces the dependence of the reconstruction on the forward spectral propagation in the wave model. Therefore, no adjustment of spectral shape is required, and the associated uncertainty is therefore avoided. The resulting dataset is therefore neither an observation-only gridded product nor an improved numerical simulation, but an offline fused SWH analysis constrained by both sources.*"

101: I'm not familiar with the CCI dataset. What is "1-Hz" here? Does this refer to one cycle per day? Is this a normal definition for "Hz"? (usually 1/sec?)

We agree that this term needed clarification for readers who are not familiar with satellite altimetry terminology. In satellite altimetry, "1-Hz" refers to a low-rate along-track product with approximately one SWH observation per second, not to one cycle per day. Many altimeter missions also provide higher-rate measurements, such as 20-Hz data, but the 1-Hz product is commonly used in operational and climate applications because averaging/resampling to this lower rate reduces random noise and improves the signal-to-noise ratio.

The text has been revised to:

"*The most important altimeter dataset used in this study is the Level-3 1-Hz significant wave height (SWH) product from the CCI (Climate Change Initiative) Sea State dataset (Dodet et al., 2020), with each record associated with its own time, latitude, and longitude. Here, "1-Hz" denotes along-track observations provided at approximately one measurement per second (corresponding to a spatial resolution of about 7 km).*"

101 again: I guess this is a daily product on a fixed, regular, lat-lon grid?

We thank the reviewer for identifying this ambiguity. The CCI Sea State product used here is not a daily product on a fixed regular latitude–longitude grid. It is an along-track altimeter product containing exact observation time, latitude, longitude, and SWH records. The revised text has been shown in the previous response.

183: Has the "parameter set" been introduced?

We agree that the term "parameter set" was introduced too abruptly. We have revised the sentence to explicitly state that the parameter set refers to the combination of the tuning parameters C, S_1, T_1, Q, S_{thr}, T_{thr}, r_{thr}. To better explain this, we have revised the corresponding part to:

"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 small overall errors was selected."

191-210: This block of text deserves a heading of some kind.

We thank the reviewer for this suggestion. We agree that the purpose of this paragraph should be made clearer. Since this block is still part of the methodological description of the fusion approach, we prefer not to introduce an additional formal subsection heading. Instead, we have added an introductory sentence before this paragraph to clarify that the following points explain the main modifications of our approach relative to conventional objective-analysis / optimal-interpolation methods:

"*With respect to the construction of the weights, while conventional objective analysis and optimal interpolation methods often define weights primarily according to spatial separation, our approach introduces several additional terms to better reflect the characteristics of wave fields:*"

Fig. 2 caption: to assist readers who are skimming, add 3-4 more words to the caption to explain what CMEMS is.

We agree and have revised the caption to briefly define CMEMS (Copernicus Marine Environment Monitoring Service). The same revision has been made in Section 2 where CMEMS dataset is first introduced.

350: for clarity, "data fusion" ==> "data fusion is applied to both datasets"

We have revised the sentence accordingly:

"***Fig. 5.*** *Comparison between WW3-ST6 (forced by ERA5 winds) and WW3-ST4 (forced by CFSR winds) before and after data fusion is applied to both hindcasts using the same altimeter observations in January 2011.*"

448: This section "Data description and availability" looks weird to me. There are two sections about data availability. Is it a normal practice for this journal to have a very small section prior to the Summary, about data availability? Why not put this content in the section after the Summary, "Code and data availability"?

We thank the reviewer for this comment. We understand the concern that the manuscript appeared to contain two data-availability sections. However, the two sections serve different purposes. The section "Data description and availability" is intended to describe the dataset produced in this study, namely the global gridded fused SWH dataset, including its two versions, file structure, variables, and metadata. This type of dataset-specific description is commonly included in dataset papers before the Summary, so that users can understand the actual product before the concluding section.

In contrast, the required "Code and data availability" section after the Acknowledgement provides access information for all data and model resources used in the study, including the CCI Sea State product, WW3-ST4/ST6 hindcasts, CMEMS gridded SWH product, NDBC buoy observations, the fused dataset, and the AI wave model of Wang and Jiang (2024), to ensure the reproducibility of this study.

468-473: The basis for these findings should be given.

We thank the reviewer for this comment. These statements were intended as methodological limitations rather than direct validation findings. First, the fusion product inherits the spatial and temporal resolution of the background wave model and the smoothing implied by the fusion procedure, so it is not designed to retain small-scale or high-frequency signals present in Level-2 along-track observations. Second, the background WW3-ST6 hindcast used in this dataset does not include wave–current interactions, so the fused product should not be used as a primary dataset for studying such interactions. We have rephrased these statements accordingly. We have revised this part to:

"*Although this high-accuracy gridded dataset offers many practical advantages, it also has some limitations compared with along-track altimeter observations. These limitations arise from the design of the fusion framework. Because the fused fields are generated on the spatial grid and temporal resolution of the numerical hindcast and involve space-time interpolation, small-scale signals present in the original along-track observations are smoothed. Therefore, the product is not intended for studies that require the full fine-scale variability of altimeter measurements, such as analyses of regional-scale SWH gradients or other small-scale wave-field features. In addition, because the WW3-ST6 background hindcast used here does not explicitly account for wave–current interactions, the fused product should not be used as a primary dataset for diagnosing such interactions, even though some wave-current-interaction signals are present in the original altimeter data. Moreover, this dataset is intended for analyses based only on total SWH, thus, other wave parameters from WW3 are not jointly adjusted within the present fusion framework, caution is warranted in applications that require strict internal spectral consistency among SWH and other wave parameters. For such applications, the consistency assumptions and associated uncertainties should be evaluated according to the objective.*"

The summary section is not too long, but is a jumble. Here's my attempted categorization of the content:

458-460 : background
460-462 : methods
462-463 : findings
463-467 : methods
468-473 : findings
474 : method
475-479 : background
479-484 : discussion
485-490 : future work

This structure means that the reader has to sort it out. If a reader just wants a bald statement of the outcome of the research, it should be easier to find.
I don't want to micromanage this, but I suggest to at least have a section or sub-section with a clear list of findings. Here, "finding" = something that is a direct outcome of the analysis and experiments described in the paper. It does not include conclusions that could have been made without the work in the paper.

We agree that the original Summary mixed background, methods, findings, limitations, and future work in a way that made the main outcomes less easy to identify. We have rewritten the Summary section to follow a clearer structure: background/motivation, method, main findings, limitations, and future work:

"Satellite altimeters can provide long-term, high-accuracy observations of SWH across the global ocean. However, their inherently sparse spatial and temporal coverage introduces undersampling errors, which can limit the reliability of wave climate analyses, particularly for extreme wave climate. To address this limitation, this study developed a global gridded SWH dataset by applying an offline fusion framework to combine WAVEWATCH III hindcast fields with along-track, jointly calibrated altimeter observations from the CCI Sea State dataset. Unlike online wave assimilation systems designed primarily for dynamically consistent reanalysis, the offline fusion framework used here aims to maximize the utilization of historical observations for retrospective SWH reconstruction. To cater to different applications, two versions of the fused dataset are provided: (1) a two-sat version, which incorporates data from only two satellites at any given time and is optimized for SWH-based wave-climate studies, ensuring broadly consistent orbital sampling characteristics throughout the record by reducing sampling-related temporal inhomogeneity; and (2) a multi-sat version, which assimilates observations from as many altimeter missions as possible and is intended for applications such as training artificial intelligence-based wave models, where high spatial and temporal fidelity is critical.

Validation against buoy measurements and withheld satellite data shows that the resulting fused gridded product achieves high accuracy, demonstrating that combining the spatial and temporal continuity of numerical hindcasts with the accuracy of altimeter observations provides a more reliable gridded SWH field than either source alone. Experiments using different background hindcasts show that the fused results are not strongly dependent on the specific numerical background field, provided that the background offers a physically reasonable estimate of the large-scale SWH distribution. Although the fused product inevitably retains some dependence on the background model, the discrepancies between products generated from different hindcasts are effectively reduced after fusion.

Although this high-accuracy gridded dataset offers many practical advantages, it also has some limitations compared with along-track altimeter observations. These limitations arise from the design of the fusion framework. Because the fused fields are generated on the spatial grid and temporal resolution of the numerical hindcast and involve space-time interpolation, small-scale signals present in the original along-track observations are smoothed. Therefore, the product is not intended for studies that require the full fine-scale variability of altimeter measurements, such as analyses of regional-scale SWH gradients or other small-scale wave-field features. In addition, because the WW3-ST6 background hindcast used here does not explicitly account for wave–current interactions, the fused product should not be used as a primary dataset for diagnosing such interactions, even though some wave-current-interaction signals are present in the original altimeter data. Moreover, this dataset is intended for analyses based only on total SWH, thus, other wave parameters from WW3 are not jointly adjusted within the present fusion framework, caution is warranted in applications that require strict internal spectral consistency among SWH and other wave parameters. For such applications, the consistency assumptions and associated uncertainties should be evaluated according to the objective.

Despite these limitations, this fused dataset already offers substantial utility across a wide range of applications. Future improvements can be made by incorporating additional satellite missions and extending the temporal coverage of the dataset. In particular, including more recent altimeter and wave-scatterometer observations, such as those from Haiyang-2 and CFOSAT, may further improve the multi-sat product and support more robust analyses of long-term wave-climate variability and trends."

Minor comments:

Fig. 2 and 3 and possibly other figures. The statistics are given to too many decimal places

Thank you for pointing this out. We have reduced the number of decimal places used for the error metrics in Figures 2 and 3 and Tables 2-4. All reported error metrics are now shown with no more than three decimal places.

Typos: Line 67, Line 68

We appreciate the reviewer for pointing out these minor issues. We have gone through the entire manuscript again to minimize typos and grammatical errors.

Response to Reviewer #2

Owing to lack of data used as a benchmark in global wave modeling authors of the current paper look at the proper approach of correction satellite altimetry products and their conclusion is quite clear. Their research focuses on the important question of how to create trustful dataset of significant wave height (Hs) by fusing multi-mission satellite altimeter observations with numerical wave models and understand its effects on precision. From technical point of view the proposed fusion method employs a space-time optimal interpolation that incorporates spatial, temporal, and Hs differences, as well as a correlation-based threshold, to optimally blend model and observational data. Furthermore, the size of the period taken into consideration is relatively small (2010-2016) to draw a reliable conclusion. The standard operational baseline period designated by the World Meteorological Organization (WMO) is 1991–2020. Besides this fact modeled wave climate is not fully Hs climate (due to wave systems partition method, e.g PTM), however multi-mission satellite wave climate is purely Hs climate (nrt, inter-calibrated, calibrated by buoys), as aptly pointed out by the authors in the main body of the text when they describe discrepancies in observations. It is not entirely clear whether we are improving the model or the observations.

I have carefully read the paper and I believe that it is an interesting study worth publishing after minor revision.

We thank the reviewer for the positive assessment of the manuscript and for raising these important points. We agree that the original description was not sufficiently precise regarding the validation period, the meaning of wave climate in the context of this study, and the respective roles of the numerical hindcast and satellite observations. We have therefore clarified these issues throughout the revised manuscript.

First, the period 2010–2016 refers only to the independent validation against the selected NDBC buoy observations. It does not represent the temporal coverage of the fused dataset or the period used to define a climatological baseline. The two-satellite fused product intended for climate-related applications covers October 1992–December 2023. The 30-year period recommended by WMO is the current standard period for calculating climatological normals. However, the purpose of the 2010-2016 analysis in this study is to quantify the accuracy of instantaneous SWH estimates against independent buoy measurements, rather than to calculate a climatological normal or derive long-term climate trends. We have revised the manuscript to make this distinction explicit and have avoided drawing climatological conclusions directly from the validation period.

Second, we agree that SWH alone does not provide a complete description of wave climate, which may also include wave period, direction, spectral shape, and partitioned wind-sea and swell components. We have therefore replaced several general references to "wave climate" with the more precise terms "SWH-based wave climate". Nevertheless, the quantities fused in the present study are physically consistent because both datasets represent total SWH.

Third, the objective of the proposed method, accurately speaking, is neither to recalibrate the original satellite observations nor to modify the physics or numerical integration of the wave model. Instead, the method constructs an observation-constrained gridded SWH dataset. The numerical hindcast provides a spatially and temporally continuous background field, while the jointly calibrated altimeter observations provide correction increments. The original altimeter measurements and the WAVEWATCH III model configuration remain unchanged. Therefore, the resulting product is therefore neither an observation-only gridded product nor an improved numerical simulation, but a fused SWH analysis constrained by both sources.

These clarifications have been added to the manuscript:

"It is noted that these buoy data are used only for independent validation and does not represent the temporal coverage of the fused dataset or a climatological reference period."

"To address the limitations outlined above, this study presents a dataset that fuses jointly calibrated satellite altimeter observations with NWM hindcast outputs. The proposed fusion framework addresses above limitations in different ways. First, it uses the jointly calibrated CCI Sea State altimeter record as the observational reference, thereby avoiding calibration against the model background. Second, because the fusion is performed offline after the hindcast has been generated, both past and future altimeter observations can be used to constrain the SWH field at a given analysis time. Third, this retrospective constraint also reduces the dependence of the reconstruction on the forward spectral propagation in the wave model. Therefore, no adjustment of spectral shape is required, and the associated uncertainty is therefore avoided. The resulting dataset is therefore neither an observation-only gridded product nor an improved numerical simulation, but an offline fused SWH analysis constrained by both sources. This dataset effectively mitigates the undersampling issue inherent to altimetry while preserving the wave variability signals captured by the altimeters, making it well suited for SWH-based wave-climate studies."

Minor comments

Lines 43-45: "...satellite based wave sensors and wave model of SAR or scatterometer..." sentencem probably need to be rephrased. Current understanding in this particular

sentence is that "wave model" can be defined in multiple meaning, e.g. Generattion/type/empirical method to retrieve. Can authors write more precisely what did they mean using term of wave model sensor.

We thank the reviewer for pointing out this ambiguous wording. The term "wave model" was a typo and should have been "wave mode," referring to the wave-imaging mode of synthetic aperture radar. We also agree that the original term "wave scatterometers" was insufficiently precise, because conventional satellite scatterometers are primarily designed to retrieve ocean surface winds rather than wave parameters. We have therefore revised the sentence to explicitly refer to synthetic aperture radar operated in wave mode and spaceborne wave spectrometers, such as CFOSAT/SWIM:

"However, satellite altimeters and other spaceborne wave-observing instruments, such as synthetic aperture radars operated in wave mode and wave spectrometers (e.g., Surface Waves Investigation and Monitoring instrument onboard China-France Oceanography Satellite), sample the ocean only along ground tracks or at spatially discrete footprints, resulting in sparse spatiotemporal coverage."

Lines 64-65 "...higher accuracy of satellite altimeter observations." Another question for auhtors: what is the measurement error according to the type and the manner by which the data is collected or stored?

We thank the reviewer for this comment. Validation of calibrated and denoised altimeter SWH products against offshore buoy measurements generally reports RMSEs of approximately 0.2 m. We have added this error information to clarify the stated accuracy of satellite altimeter observations:

"A more promising approach for generating a high-quality SWH dataset is to combine the seamless spatiotemporal coverage of NWM hindcasts with the higher accuracy of satellite altimeter observations, which generally have an RMSE of approximately 0.2 m when validated against offshore buoy measurements (Jiang, 2023)."

Lines 67:70: Besides the fact that ERA5 wave model assimilation of along track data with optimum interpoaltion method (of 5 satellite missions out of 10 that are in operational use nowdays) there is WAVERYS (MERCATOR) https://data.marine.copernicus.eu/product/GLOBAL_MULTIYEAR_WAV_001_032/description that also assimilates satellite altimetry products (such as family of missions named after Sentinel and SAR). Do you agree with the statetment that for the purpose of validation numerical models versus observations: already assimilated

satellite measurements could not be used for postrpocessing output as it consists of data fusion (model+observations)?

Wave reanalysis products such as ERA5 and WAVERYS, which have already assimilated satellite altimeter observations, are not suitable as background fields for the present offline fusion with altimeter data. Otherwise, the same or closely related observational information would be used twice, making it difficult to isolate the contribution of the proposed fusion and potentially introducing dependence between the background field and the observations. For this reason, we use a WAVEWATCH III hindcast that does not assimilate wave observations as the background field. We have clarified this point and added WAVERYS as another example of an assimilative wave reanalysis:

"Another example of wave reanalysis dataset is the WAVERYS, which also assimilates SWH observations from historical altimeter missions (Law-Chune et al. 2021). However, the "online" assimilation methodology in ERA5 and WAVERYS presents several limitations when used for historical SWH data reconstruction"

"Assimilative wave reanalyses such as ERA5 are not suitable as background fields for the present study, because subsequent fusion with altimeter observations would lead to repeated use of the same observational information. Therefore, the SWH data from a WAVEWATCH III (WW3) hindcast based on the source term package 6 (ST6), without wave data assimilation, is adopted as the background field."

Fig 4, 6 should be rearanged and colormaps should be enhanced especially to repeat the same colormap for RMSE, CC. Bias on Figs. 4-7 is not meaningful, better enhance the limits.

We thank the reviewer for this suggestion. We assume that "Figs. 4 and 6" refers to Figs. 4 and 5, because Fig. 6 does not contain RMSE or CC panels. Within each figure, identical colorbar limits are used for the corresponding panels to allow direct comparison before and after fusion. However, we think the colorbar limits are not necessarily identical between Figs. 4 and 5 because the two figures represent different comparisons. Figure 4 evaluates the original and fused fields against independent altimeter observations, whereas Fig. 5 compares results obtained using two different numerical hindcasts as background fields. The magnitudes and distributions of the corresponding errors are therefore different. Using limits adapted to the data range in each figure provides better visual contrast and more clearly reveals the spatial error patterns. The same colormap type is retained for each metric to ensure visual consistency.

We are also not entirely certain how to interpret the statement that “Bias on Figs. 4–7 is not meaningful” and the suggestion to “enhance the limits”. We respectfully consider the bias information in these figures to be meaningful for different purposes. In Fig. 4, the comparison of bias before and after fusion demonstrates the substantial reduction in systematic errors achieved by the proposed method. In Fig. 5, the bias between the results obtained using two different background hindcasts is used to evaluate the sensitivity of the fused product to the choice of background field. In Fig. 6, the differences between the fused and observation-only gridded products illustrate that satellite undersampling can introduce substantial systematic deviations. In Fig. 7, bias is reported together with the other standard error metrics and confirms that the remaining systematic error is small.

We have therefore retained the bias results in these figures. We would appreciate further clarification if our interpretation of the reviewer’s comment is not correct.

Response to Reviewer #3

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.

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-of-freedom, 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. 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_l = 30, 90, 180 min; Q = 2000, 4000, 8000, 16000; S_{thr} = 1000, 2000, 4000 km; T_{thr} = 6, 12, 24, 48 h; and r_{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 do not 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 empirical 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):*

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

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

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

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 have clarified 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 (or even impossible) 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 have added a new subsection, Section 3.3 "Temporal homogeneity test for mission transitions", to examine this issue more explicitly. In the revised manuscript, we added a new figure showing time series of monthly mean SWH from the WW3-ST6 hindcast background, the "two-sat" fused dataset, and the fused-minus-background difference during 1992–2023. The analysis is conducted for the global ocean and for several latitude bands, including low-, mid-, and high-latitude regions in both hemispheres. The major satellite-transition times in the two-satellite configuration are marked by vertical dashed lines (as shown below).

The purpose of this analysis is to diagnose whether satellite replacements introduce abrupt step-like changes in either the fused SWH or the observation-derived correction relative to the background field. The results show that the fused SWH generally follows the temporal variability of the WW3-ST6 background, including the strong seasonal cycle in the mid- and high-latitude oceans. The fused-minus-background difference also exhibits clear regional and seasonal variations, but these variations evolve smoothly across the major mission-transition times. No obvious discontinuity synchronized with the satellite replacements is observed in either the fused SWH or the difference time series. The added subsection reads:

“3.3 Diagnosis of mission-transition discontinuities

The purpose of the “two-sat” version is to provide a dataset with improved temporal homogeneity for wave climate studies. However, maintaining a constant number of fused altimeters does not by itself guarantee the homogeneity of the time series. The replacement of satellite missions may still introduce artificial discontinuities through differences in orbit, sampling geometry, instrument characteristics, or residual inter-mission calibration uncertainties.

Therefore, we analyzed the possible discontinuities associated with satellite mission transitions in the “two-sat” product. Fig. 5 shows the time series of monthly mean SWH from the WW3-ST6 hindcast background, the “two-sat” fused dataset, and the fused-minus-background difference during 1992–2023 for the global ocean and different latitude bands. The fused SWH generally follows the temporal variability of the WW3-ST6 background, including the pronounced seasonal cycle in the mid- and high-latitude bands, while applying a relatively stable positive or negative correction depending on region. The fused-minus-background difference also exhibits clear regional and seasonal characteristics, reflecting systematic differences between the altimeter-constrained analysis and the numerical background.

Importantly, these differences evolve smoothly across the major satellite-transition times marked by the vertical dashed lines. Although the magnitude and seasonality of fused-minus-background difference vary among latitude bands, particularly in the mid- and high-latitude oceans where wave variability is stronger, no abrupt shift appears to be synchronized with the replacement of satellite missions. This indicates that the two-satellite configuration does not introduce visually detectable step-like discontinuities in the monthly mean SWH record. Therefore, no obvious discontinuity synchronized with the mission transitions is observed in either the fused SWH or the difference time series, supporting the temporal homogeneity of the two-satellite product across mission transitions.

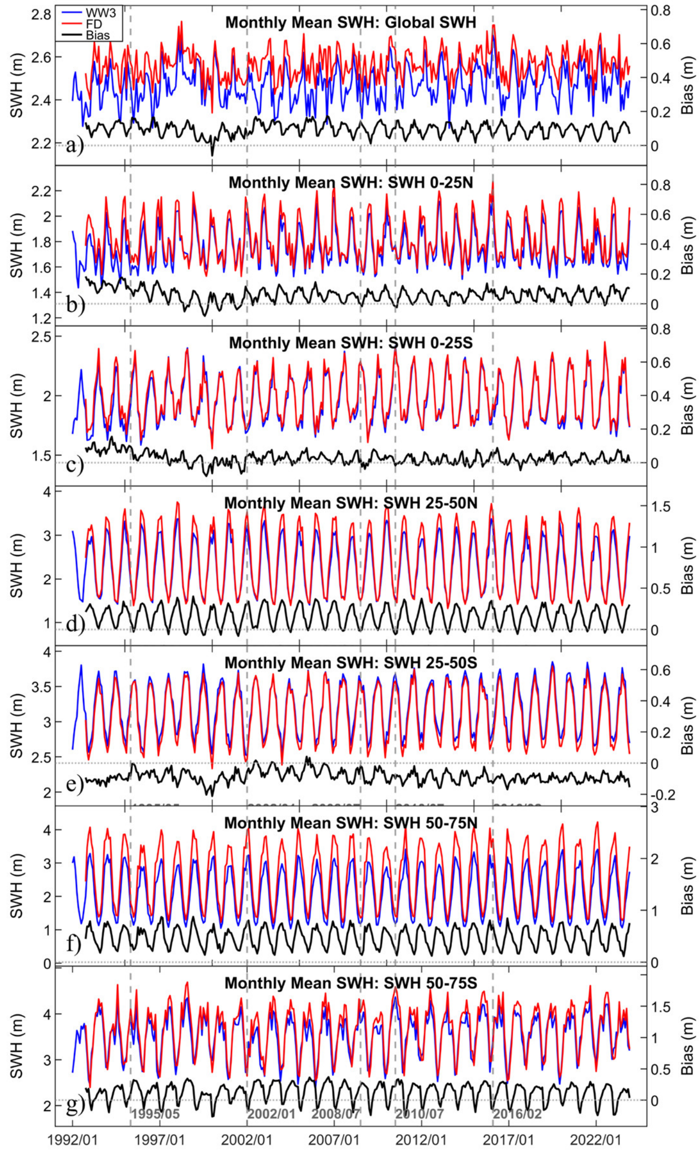

Fig. 5. *Time series of monthly mean SWH from the WW3-ST6 hindcast background (blue), the "two-sat" fused dataset (red), and fused-minus-model differences (black) during 1992–2023: (a) the global ocean, (b) 0–25°N, (c) 0–25°S, (d) 25–50°N, (e) 25–50°S, (f) 50–75°N, and (g) 50–75°S. Vertical dashed lines indicate the major satellite-transition times in the two-satellite configuration, as shown in Table 1."*

Regarding the reviewer's suggestion to provide a global map of SWH trends, we agree that trend analysis is an important application of the dataset. However, we prefer not to include a full global trend assessment in the present data-description paper, because robust trend estimation requires dedicated treatment of trend significance, serial correlation, regional climate variability, and uncertainty propagation, and is therefore better addressed in a dedicated wave-climate analysis paper. Moreover, a trend map is not the most direct diagnostic for the specific issue raised here, namely possible discontinuities at satellite mission transitions. We therefore focus the revised analysis on transition-centred time-series diagnostics, which more directly address whether the mission replacements introduce artificial discontinuities in the "two-sat" product.

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 mainly "to reduce the volume of data processing" was not sufficiently clear. It may have given the impression that the validation year was selected only for convenience, which was not our intention. We have revised the manuscript to better clarify the rationale for using 2011 as the intensive satellite-validation year and to distinguish methodological validation from long-term homogeneity assessment.

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 have therefore revised the wording to clarify that the 2011 experiment is a methodological validation case enabled by the availability of multiple simultaneous altimeters. We have also added the mission-transition consistency analysis described in response to the previous comment.

"We further evaluated the global performance of the fused dataset using altimeter observations from 2011. This year was selected because four altimeter missions were available simultaneously, enabling multiple leave-one-satellite-out and two-satellite-combination experiments. In each experiment, one or more satellite missions were excluded from the fusion and then used as withheld references for evaluation."

"Tables 2–4 summarize the evaluation results of bias, RMSE, and CC, respectively. In each experiment, two of the four available satellites were fused with WW3-ST6, while the remaining satellites were used as independent references. It is noted that these experiments are intended to evaluate the stability and achievable accuracy of the fusion method under different observational configurations, rather than to provide a complete year-by-year assessment of the full record, thus, only one year of data is used."

"These results demonstrate the robustness and accuracy of the fusion method under different satellite-combination configurations. In particular, the method consistently reduces bias and RMSE and increases CC relative to the original WW3-ST6 background, regardless of which pair of satellites is used for fusion and which mission is used as the withheld reference. Since the final fused dataset is generated using the same fusion framework, the same calibrated CCI Sea State observations, and the same WW3-ST6 background, these results provide strong evidence supporting the robustness and accuracy of the resulting dataset."

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 might not be sufficient for a data-description paper, and that the manuscript itself should provide a concise and complete overview of the released products.

We have added a table in the revised manuscript summarizing the key properties of the data product. This table includes the exact temporal coverage, spatial resolution, temporal resolution, available variables, missing-value convention, background WW3-ST6 information, and approximate total data volume.

The corresponding text related to the table is now:

"The main characteristics of the released fused SWH products are summarized in Table 5. Meanwhile, each NetCDF file corresponds to one calendar month, with global attributes documenting metadata such as the source WW3-ST6 hindcast file, the names of the fused satellites, the start and end time, the identifiers of the fused satellites, and the percentage contribution of each satellite dataset to the fusion."

Table 5. Main characteristics of the released fused SWH products.

Information	Description
Product version	Multi-sat / Two-sat (with the same format)
Time period	October 1992 – December 2023 *
Data coverage	Global
Spatial resolution	0.5°×0.5°
Temporal resolution	3 hours
Missing-value convention	NaN for land, masked coastal regions, and sea-ice-covered grid cells.
Single file volume	~ 480 MB (for one month per version)
Total dataset volume	~ 360 GB + 360 GB
Version of WW3 hindcast	v7.00, ST6 source-term package, ERA5 10-m wind forcing, 0.25° × 0.25°, 3 h.
Altimeter data fused	CCI Sea State v4, Level-3 1-Hz along-track SWH, Daily Multi-Sensor product.
Available variables	Fused SWH value (m), Distance-related parameter indicating the proximity to the nearest altimeter constraint (R value), Latitude, Longitude, Time

* The CCI Sea State v4 record starts in August 1991, but the released fused products start in October 1992 because the two-sat configuration requires two simultaneously available altimeter missions.

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 have therefore revised this wording to state explicitly that the released fused product has a native resolution of 0.5° × 0.5°:

"The fused dataset effectively mitigates these undersampling issues. In particular, it provides spatially complete SWH fields on a regular 0.5° × 0.5° grid, thereby reducing the need to account for the strong grid-size dependence that arises when sparse along-track observations are directly gridded."

We have also clarified 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 released fused products start later because the two-sat product requires a suitable pair of altimeter missions satisfying the prescribed sampling configuration. Before October 1992, only one altimeter mission is available in the CCI record used here, so the prescribed two-satellite configuration cannot be constructed. For consistency between the two released product versions, both the two-sat and multi-sat products are therefore provided from October 1992 onward. This distinction is now stated explicitly in the revised version:

"Although the CCI Sea State v4 record starts in August 1991, the fused product starts in October 1992 because, before this time, only one altimeter mission is available in the CCI record and the prescribed two-satellite configuration cannot be constructed. For consistency, the multi-sat product is released over the same temporal period as the two-sat product."

In addition, we briefly address this issue in the footnote of Table 5, as shown above. 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 have revised the Methods section to state more clearly how sea-ice-affected regions are treated in the current processing. We have also added a caution 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.

"It is noted, however, that no dedicated sea-ice-specific correction, dynamic ice-edge filtering scheme, or coastal-wave correction is applied in the present fusion procedure. Sea-ice-affected regions are treated according to the availability and quality control of the input WW3 and CCI Sea State data, together with the general observation-screening criteria described above. Additional uncertainties are therefore expected near sea-ice margins, where the WW3 background may be less accurate because of ice-wave interactions and possible ice-mask uncertainties, and where valid altimeter observations are generally sparser and may provide weaker local constraints. Similarly, although altimeter observations within 75 km of coastlines are excluded to reduce nearshore representativeness errors, fused SWH values near coastal regions should also be interpreted with caution because no dedicated coastal-wave correction is applied. Therefore, the product is primarily intended for open-ocean wave applications, and values near coastlines and sea-ice margins should be used with caution."

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.

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

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

We thank the reviewer for pointing this out. The NDBC wave records used in this study are reported at hourly time stamps, while the reported SWH is calculated from the waves observed during a 20-minute sampling period. Therefore, the “20-min averaged SWH” refers to the internal sampling interval used to derive each hourly reported SWH value, rather than to a 20-minute temporal resolution of the buoy dataset. We have revised the data introduction to make it clear:

“The hourly SWH records derived from 20-min wave sampling periods obtained from the National Data Buoy Center (NDBC) are used as a reference in this study.”

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

This table has been added to the manuscript and has been shown in above replies.

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

We thank the reviewer for this comment. We would like to clarify that the number of valid matchups is already displayed in the scatter plots for the main validation comparisons, including the buoy comparison and the Jason-2 leave-one-satellite-out validation. We may therefore have misunderstood the reviewer’s request. Could the reviewer clarify whether this comment refers specifically to the satellite-combination experiments summarized in Tables 2–4? In Tables 2–4, all validation experiments are based on along-track altimeter observations from the same year, with two missions used for fusion and one mission used as reference datasets. Therefore, although the exact number of valid collocations may differ slightly among missions because of orbital sampling, data availability, and quality control, the sample sizes are expected to be broadly comparable and not to differ by orders of magnitude. The interpretation of the relative performance among these experiments is therefore not expected to be sensitive to sample-size differences.

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, exceeding 8,000,000 in this comparison, the difference is expected to be statistically detectable.

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 have revised 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:

"The inclusion of the modelled-SWH difference term reduces the global RMSE only slightly, by approximately 0.005 m, and this small reduction should not be over-interpreted as a major improvement in the domain-mean error metric. The term is used mainly because it improves the physical consistency of local correction increments under model-observation displacement conditions. When a modelled storm or swell system has an arrival-time or position error, geographically or temporally nearby observations may correspond to dynamically different sea states. The SWH-based distance helps reduce such inappropriate influences by assigning larger weights to observations with similar modelled SWH and smaller weights to observations that are likely outside the wave system."

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 have reorganized the Code and Data Availability section to clearly separate, which now reads:
"Input datasets used for fusion: The CCI sea state product is available at https://data-cersat.ifremer.fr/data/ocean-waves/cci-seastate/v4/. WAVEWATCH III ST4 and ST6 hindcasts are available from Rascle and Ardhuin (2013) and Liu et al. (2021), respectively. Reference datasets used for validation and comparison: The gridded multi-mission merged satellite SWH data product from CMEMS is available from https://doi.org/10.48670/moi-00180 (CMEMS 2024). The NDBC buoy data are available from https://www.ndbc.noaa.gov/. The AI wave model is available from Wang and Jiang (2024). The final fused products released: The fused dataset is available at https://doi.org/10.57760/sciencedb.29314 (Su and Jiang 2025)."

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 have revised the figure caption to state the time period explicitly:

*"**Fig. 4.** Spatial distributions of SWH error metrics between WW3-ST6 and Jason-2 (a-c) before and (d-f) after fusing SWH data from CCI-Sea State in 2011: (a & d) bias, (b & e) RMSE, and (c & f) CC."*

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 have revised 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):

"Figure 7 is reproduced from an independent-test-year results of Wang and Jiang (2024), in which 2020 was withheld from model training and used exclusively for evaluation."

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.

Response to Reviewer #4 (Informal reviewer)

While the authors present a high-resolution gridded product, the scientific justification is weak. By bypassing spectral information, the product lacks the physical consistency. Furthermore, given that established reanalyses (ERA5) already assimilate this data at the spectral level, and existing climate-standard products (ESA CCI, Ribal/Young) already address temporal consistency, this manuscript offers a redundant statistical patch rather than a significant scientific advancement. The risk of promoting circularity by using this product to train future AI models outweighs the convenience of having a pre-fused grid.

We thank the reviewer for the thorough and critical evaluation of our manuscript. The comments have helped us clarify the motivation, scope, and limitations of the proposed dataset. However, several of the criticisms appear to be based on assumptions that differ from what is actually described in the manuscript. In particular, the proposed fusion method does not use AI, and the released product is not presented as a corrected full-spectrum wave reanalysis.

We would like to first clarify the fundamental objective of this study. The proposed product is not intended to replace physically based wave re-analyses, modify the underlying numerical wave model, or provide a corrected full-spectrum wave reanalysis. Instead, it aims to construct a globally complete, hourly gridded SWH dataset by combining the complementary advantages of satellite altimeter observations and numerical hindcasts.

Satellite altimetry provides highly accurate local SWH observations but suffers from severe spatiotemporal undersampling. Conversely, numerical wave models provide physically consistent and spatially complete fields but contain model-dependent uncertainties. The purpose of the proposed offline fusion is therefore to reduce the sampling limitations of satellite observations while constraining the model background toward observations. The resulting dataset should be interpreted as an observation-constrained gridded SWH analysis, rather than an improved along-track satellite product or an improved numerical wave model.

We have revised the manuscript to clarify this scope and have added additional discussion regarding the limitations of SWH-only correction, the relationship with existing reanalysis products, and appropriate applications of the dataset.

1. Lack novelty (ERA5 redundancy)

The primary scientific justification for a new data product is that it fills a gap or offers a significant improvement over existing, established datasets.

->ERA5 (and the upcoming ERA6) already assimilates satellite altimetry (using 4D-Var or similar frameworks) to produce a globally gridded SWH product.

->If the authors' "fusion" is simply a statistical post-processing of a hindcast, it is effectively a Reanalysis-lite. The manuscript must demonstrate that their product significantly outperforms ERA5-Wave in a way that isn't just fitting the observations closer (which is trivial with ML). If the improvement is marginal, the community is better served by the physically consistent ERA5 than by a fragmented, offline statistical product.

We thank the reviewer for this important comment. We fully agree that ERA5 and other wave re-analyses represent major achievements in global wave reconstruction and provide valuable, dynamically consistent wave products. However, we respectfully disagree with the premise that the present dataset is redundant unless it universally outperforms ERA5. The two products are designed for different scientific objectives and involve substantially different treatments of satellite observations.

First, we would like to clarify the data-assimilation method used for ocean waves in ERA5. Although the atmospheric component of ERA5 is produced using 4D-Var, altimeter SWH observations in the wave component are assimilated using a sequential optimal interpolation (OI) scheme, rather than 4D-Var. Therefore, the reviewer's reference to variational method does not apply to the assimilation of altimeter SWH observations.

Second, the proposed dataset is not generated using machine learning, and no AI-based fitting is involved at any stage. It is produced using an explicitly defined space-time fusion method that combines a non-assimilative WW3 hindcast background with jointly calibrated multi-mission altimeter observations. Consequently, the characterization of the method as trivially fitting observations using ML is not applicable to the present study.

Third, ERA5 and the proposed dataset have different objectives. ERA5 aims to generate a dynamically consistent wave reanalysis through sequential model integration and assimilation. The proposed dataset instead aims to produce a retrospective, observation-constrained gridded SWH reconstruction. The offline framework can use observations located both before and after the target time within the specified fusion window. In contrast, the sequential OI system does not retrospectively revisit an already produced historical state using observations from later assimilation windows. This difference is important for historical reconstruction because the offline method can more fully exploit the available satellite sampling around each target time, although it is clearly not intended for operational forecasting.

Fourth, the two approaches treat satellite biases differently. In the ECMWF wave-assimilation system, bias correction is applied to harmonize altimeter SWH observations with their model counterparts before the OI analysis. This procedure is appropriate for maintaining a stable and dynamically consistent assimilative system. However, because the correction is defined relative to the model background, it may, in principle, attenuate observational signals that systematically differ from the model baseline. In the present study, the altimeter observations are jointly calibrated across satellite missions independently of the WW3 background and are not regressed toward the background model climatology before fusion. This distinction is particularly relevant when constructing a long-term SWH record intended to retain observational variability.

We therefore do not consider a universal reduction in errors relative to ERA5 to be the sole criterion for the scientific value of the dataset. ERA5 prioritizes dynamical and spectral consistency, whereas the present product prioritizes the direct incorporation of jointly calibrated altimeter SWH information into a spatially and temporally continuous gridded record. These are complementary rather than competing objectives.

Accordingly, the proposed dataset should not be interpreted as a simplified replacement for ERA5 or as a “Reanalysis-lite”. It is an alternative observation-constrained SWH product that bridges the gap between accurate but sparsely sampled along-track altimeter observations and continuous but model-dependent numerical hindcasts.

2. Physical inconsistency

The authors explicitly state that their method bypasses the need for wave spectral information. To expert users, this might be a major issue.

->SWH is an integrated parameter of the 2D wave spectrum
->By correcting only the SWH using a black box statistical method, the authors create a physical mismatch between SWH and all other spectral parameters (Mean Period, Peak Direction, etc.). A dataset where the SWH is "corrected" but the underlying spectrum remains tied to the original erroneous hindcast could be dangerous… users might perform secondary analyses (like wave power or stress calculations) that are physically impossible. There is not much discussion about this or analysis to show that they thought it through.

Yes, SWH is an integral parameter and a corrected SWH should not be arbitrarily combined with uncorrected spectral parameters and interpreted as a physically self-consistent wave state. However, we would like to clarify two important points regarding the present dataset.

First, the proposed method is not a black-box statistical correction. It is an explicitly formulated space-time weighted fusion method, and no machine-learning or AI model is involved in generating the dataset at all.

Second, the released product contains only total SWH. It does not provide a corrected directional spectrum, wave period, wave direction, or spectral partitions. Therefore, the dataset itself does not contain an internally inconsistent combination of corrected SWH and uncorrected spectral variables. Such an inconsistency would arise only if a user combined the fused SWH with spectral parameters from the original hindcast and interpreted them as a jointly corrected wave state, which is outside the intended scope of the product.

The decision not to modify the spectrum is deliberate. Altimeter observations provide total SWH but do not uniquely determine the underlying two-dimensional wave spectrum. Reconstructing or adjusting the full spectrum from SWH alone is an underdetermined problem and necessarily requires additional assumptions that may introduce further uncertainty.

In principle, the fused SWH field could subsequently be used to adjust the model spectrum using an approach similar to that adopted in the ECMWF wave assimilation system. In that system, an analyzed SWH field is first obtained through optimal interpolation, after which the background spectrum is rescaled in energy and frequency using different assumptions for wind-sea and swell components. However, such spectral adjustment requires additional assumptions regarding spectral shape and wave-system composition and may introduce uncertainties beyond those associated with the SWH analysis itself. Developing and evaluating a physically consistent spectral correction is therefore beyond the scope of the present SWH dataset.

We nevertheless agree that the intended scope of the product and the implications for multi-parameter applications should be stated more clearly. We have therefore added a caution in the revised manuscript. The revised text clarifies that the fused SWH may be used jointly with wave parameters from other products, as is common in multi-source analyses, but that these variables have not been jointly adjusted within the present fusion framework. Users should therefore evaluate the relevant consistency assumptions and uncertainties when their applications require strict internal consistency among SWH, wave period, wave direction, and the underlying directional spectrum:

"Moreover, this dataset is intended for analyses based only on total SWH, thus, other wave parameters from WW3 are not jointly adjusted within the present fusion framework, caution is warranted in applications that require strict internal spectral consistency among SWH and other wave parameters. For such applications, the consistency assumptions and associated uncertainties should be evaluated according to the objective."

3. Circularity

One of the stated goals is to provide a multi-sat version for training future AI wave models.

->ML should be trained on the best possible representation of reality (observations) or high-fidelity physics

->If we use an AI-corrected product to train other AI models, we are creating an endless loop. We risk training future models on the artifacts and biases of the authors' specific statistical correction rather than on true ocean physics. This data incest can lead to overconfident models that fail when faced with non-linear extreme events that were not in the training fusion logic (they are smoothed out or highly distorted).

We thank the reviewer for raising the general concern that uncertainties in a training dataset may be inherited by downstream AI wave models. However, the concern expressed here is based on an incorrect characterization of the proposed product as an "AI-corrected product". We reiterate that no AI method is used at any stage in generating the fused dataset. The product is constructed using an explicitly defined space-time optimal interpolation framework that combines jointly calibrated altimeter observations with a non-assimilative WW3 hindcast. Therefore, using this dataset to train an AI wave model does not constitute an AI-to-AI feedback loop or the recursive use of an AI-generated product.

More fundamentally, even if an AI method were used for the data-fusion step, which is not the case in the present study, this would not by itself constitute an AI-to-AI feedback loop. An observation-fusion model and an AI wave model solve fundamentally different problems. A fusion model estimates an analysis field from sparse observations and a background field, whereas an AI wave model predicts the wave field from atmospheric forcing and, depending on the model formulation, sometimes an initial wave state. Their inputs, objectives, and learned mappings are therefore different. The use of an observation-constrained analysis as a target for a prognostic AI model is not inherently circular merely because both procedures may employ AI.

The more relevant scientific concern is instead whether residual errors or artifacts in the fused product may be inherited by the downstream model. We agree that this possibility must be considered. However, the same issue applies to any AI model trained using reanalysis, analysis, remotely sensed Level-4 products, or numerical hindcasts; none of these products constitutes an error-free representation of reality.

Training AI wave models often requires spatially and temporally complete target fields. Direct along-track altimeter observations are highly valuable but are too sparse to provide complete target fields for supervised global wave-field modeling. The fused dataset is intended to bridge this gap by providing an observation-constrained gridded SWH field. This is conceptually similar to the widespread use of reanalysis or analysis products as training targets, although the present product is more directly constrained by jointly calibrated altimeter observations. The supplied distance-related variable R also indicates the proximity of each grid point to the nearest altimeter observation and can be used to assign greater training weight to samples that are more directly constrained by observations.

We acknowledge that the fused dataset should not be regarded as an error-free representation of reality. Like any observation-constrained analysis or reanalysis, it may retain residual uncertainties associated with the background field, observational sampling, and the assumptions of the fusion method. AI models trained using this product should therefore be independently evaluated against withheld satellite and buoy observations, particularly for extreme sea states and conditions that are poorly represented in the training record.

4. Time-space issues

The authors mention that their offline approach allows retrospective correction of past model outputs using future observations.

->While mathematically sound for gap-filling, this is physically problematic for wave climate analysis.

->Waves propagate… using future satellite passes to correct a past state in a non-causal way can smear out the temporal evolution of swell events. For wave climate trends (their two-sat version), this smoothing might hide the very variability or subtleties the researchers are looking for…

We thank the reviewer for raising the concern regarding the preservation of the temporal evolution of swell events. However, we respectfully disagree that the use of observations acquired after the target time is inherently non-physical or necessarily causes temporal smoothing in an offline historical reconstruction. Temporal smoothing may occur when observations representing different stages of an evolving event are blended using excessively broad or insufficiently constrained weights. The relevant issue is therefore how the observational influence is formulated, rather than whether an observation is acquired before or after the target time.

The use of later observations to constrain an earlier state is well established in retrospective estimation. For example, 4D-Var adjusts the model state at the beginning of an assimilation window so that the resulting trajectory is consistent with observations distributed throughout that window. Consequently, observations acquired later in the window can constrain an earlier model state through the model dynamics. We do not suggest that the present method is equivalent to 4D-Var. This example simply demonstrates that the use of later observations is not inherently physically invalid when their influence is appropriately constrained.

In the present method, temporal separation is explicitly incorporated into the distance metric. Observational influence progressively decreases with increasing temporal separation, and observations separated from the target time by more than 24 h are excluded. The influence is further constrained by spatial separation, the difference in modelled SWH between the target and observation locations, and the correlation between their background SWH time series. These constraints are intended to reduce the contribution of observations representing unrelated wave systems or substantially different stages of the same event.

Moreover, the dynamically evolving WW3 hindcast provides the background spatial and temporal structure, including the modelled generation, propagation, and decay of swell systems. The altimeter observations provide constrained correction increments to this evolving background rather than replacing its temporal evolution through direct interpolation of satellite measurements. Therefore, the potential for smoothing would arise if temporally distant or dynamically inconsistent observations were assigned excessive weights, which the present formulation is specifically designed to avoid.

Finally, the fused product is evaluated against instantaneous independent buoy measurements and withheld altimeter observations, rather than only against monthly means or climatological statistics. These independent evaluations show that the fusion improves instantaneous SWH accuracy relative to the original hindcast. The results therefore do not support the characterization of the method as producing an excessively smoothed reconstruction of evolving wave fields.

5. Climate trend justification

The authors claim their "two-sat" version is for climate studies to ensure temporal consistency.

->The ESA CCI Sea State dataset already exists for this exact purpose, using a very rigorous, inter-calibrated multi-mission approach.

->The manuscript should address: Why is this better than the ESA CCI product? If the answer is just "it's on a grid," that is a processing step, not a significant contribution… Gridding can be done by anyone with a basic interpolation script; it doesn't necessitate a new global dataset publication.

We fully agree that the ESA CCI Sea State dataset is a rigorous, inter-calibrated multi-mission observational record developed specifically to support climate applications. Indeed, the present fused dataset is built directly upon the CCI Sea State observations and should be regarded as a complementary product rather than a replacement for the original CCI record.

The term "temporal consistency" was not sufficiently defined in the original manuscript. In the CCI Sea State dataset, temporal consistency primarily refers to the calibration and harmonization of measurements from different satellite missions. In the present two-satellite product, it refers instead to the consistency of the sampling configuration and observational influence used in the reconstruction. Inter-calibration can reduce systematic differences between missions, but it cannot eliminate the temporal inhomogeneity caused by changes in the number of operating satellites, their orbital configurations, and the associated probability of sampling individual storms and extreme sea states. This distinction is particularly relevant for extreme-wave statistics. As the number of concurrent altimeters increases over time, the probability that short-lived extreme events are sampled also increases. Consequently, even a perfectly inter-calibrated observational record may exhibit time-dependent sampling characteristics. The effects of this temporally uneven sampling have been systematically investigated in a previous study by the first author (Jiang 2020).

The two-satellite configuration is designed to reduce this source of inhomogeneity by maintaining a relatively stable number of observations and a broadly consistent combination of orbital sampling patterns throughout the record. Specifically, one approximately 66°-inclination mission is paired with one near-polar mission, with replacements selected from satellites having similar orbital characteristics when necessary.

The two-satellite configuration is designed to reduce this source of inhomogeneity by maintaining a relatively stable number of observations and broadly consistent orbital sampling characteristics throughout the record. Specifically, one mission with an orbital inclination of approximately 66° is paired with one near-polar mission, and replacements are preferentially selected from satellites with similar orbital characteristics when necessary. We agree, however, that the original wording "ensuring temporal consistency" was too strong. The two-satellite configuration cannot guarantee that the resulting dataset is free from all temporal inhomogeneities or that estimated climate trends are unbiased. Residual uncertainties associated with cross-mission calibration, observational noise, and the numerical background remain. We have therefore revised the wording to state that the two-satellite product is designed to reduce sampling-related temporal inhomogeneity:

Abstract:

"A 'two-sat' version that incorporates data from only two satellites at any given time, designed for wave climate studies. This configuration is designed to reduce sampling-related temporal inhomogeneity by maintaining a relatively stable number of observations and broadly consistent orbital sampling characteristics throughout the record."

Section 2.3 Two Types of Fused Data:

"The 'multi-sat' product prioritizes accuracy by maximizing observational coverage, whereas the 'two-sat' product is designed to reduce sampling-related temporal inhomogeneity by maintaining a relatively stable number of observations and broadly consistent orbital sampling characteristics. Here, sampling consistency refers to reducing temporal changes in observation density and orbital sampling characteristics and is distinct from the cross-mission calibration consistency already provided by the CCI Sea State dataset."

"For the 'two-sat' product, the selection of satellites is guided by sampling characteristics: one satellite with an orbital inclination of approximately 66° is paired with another in a near-polar orbit to maintain broadly consistent orbital sampling characteristics over time"

Summary:

"… a 'two-sat' version, which incorporates data from only two satellites at any given time and is optimized for SWH-based wave-climate studies, ensuring broadly consistent orbital sampling characteristics throughout the record by reducing sampling-related temporal inhomogeneity"

We also agree that simply converting along-track observations onto a grid would not, by itself, constitute the principal scientific contribution of this study. However, the proposed dataset is not produced through basic geometric interpolation or coordinate remapping. The central difficulty is not the grid format, but the reconstruction of wave conditions that were never sampled by an altimeter. Because ocean waves vary rapidly in both space and time, direct interpolation of sparse satellite tracks cannot recover storms or swell systems that fall between tracks or between satellite overpasses. Increasing the sophistication of the interpolation algorithm does not create the missing observational information.

The present method therefore uses a dynamically evolving WW3 hindcast as a spatially and temporally complete background field and applies observation-derived SWH increments through a constrained space-time optimal interpolation framework. The observational weights depend not only on geographical distance, but also on temporal separation, modelled-SWH similarity, and the correlation between the target and observation locations. The resulting product is therefore a model-assisted, observation-constrained reconstruction rather than a conventional gridding of the CCI observations.

The importance of this distinction is supported by the validation results. In the buoy comparison, the altimeter-only gridded product considered in the manuscript had an RMSE of 0.37 m, which was higher than the 0.31 m RMSE of the non-assimilative WW3 hindcast, while the fused dataset had an RMSE of 0.27 m and performed better for extreme cases. These results show that gridding sparse altimeter observations alone does not necessarily produce a more accurate spatially complete SWH field.

We also recognize that CCI provides a monthly Level-4 gridded product, and we have clarified this more explicitly in the manuscript. The CCI Level-4 product and the present fused dataset nevertheless remain fundamentally different. The former represents statistics derived from the satellite samples available within each grid cell and time interval, whereas the latter provides a spatially complete time series constrained by both the numerical background and the satellite observations. In the comparison presented in the manuscript, residual ground-track patterns remain visible in the monthly CCI Level-4 field, and the differences become particularly pronounced for upper-percentile SWH because of the limited probability of sampling individual storms

The CCI Sea State record remains the fundamental observational source of this dataset, and is more appropriate when direct measurement traceability and along-track variability are required. The added value of the present product is that it uses these high-quality CCI observations to construct and independently validate a spatially complete, temporally resolved SWH reconstruction. The two-satellite version specifically aims to reduce changes in the degree and geometry of observational constraint over the multi-decadal record.

Ref:
Jiang, H. (2020). Evaluation of altimeter undersampling in estimating global wind and wave climate using virtual observation. *Remote Sensing of Environment*, *245*, 111840.

6. More broadly...

->Scientific progress comes from fixing the source terms (Sin, Sds) in models like WAVEWATCH3.

->This work treats the model as a "black box" that is "wrong" and needs a statistical patch. In a data journal, the goal should be to provide data that reveals new truths. If the corrected data simply masks model bias without explaining it, it serves as a band-aid that discourages researchers from doing the hard work of improving model physics.

We fully agree that the continued development of source terms, including wind-input and wave-dissipation parameterizations, is important for advancing numerical wave modelling. However, we respectfully disagree with the premise that progress in wave science and engineering can arise only from fixing the source terms of numerical wave models, or that observation-model fusion and data assimilation are inherently opposed to improvements in model physics.

Numerical model development, satellite observations, in situ measurements, data assimilation, data reconstruction, and model evaluation are complementary rather than competing scientific activities. In particular, reliable observational datasets are indispensable for identifying systematic model deficiencies, evaluating alternative physical parameterizations, and determining whether a proposed change actually improves model performance. Improving model physics and producing more accurate reconstructions of the historical ocean state therefore address different scientific objectives and can support one another.

The characterization of WW3 as a “black box” or a “wrong” model in the present study is incorrect. The WW3 hindcast is used precisely because its physics-based integration provides a spatially and temporally complete representation of the modelled generation, propagation, and dissipation of wave systems. The manuscript explicitly describes the source-term configuration, atmospheric forcing, spatial and temporal resolution, and validation accuracy of the WW3-ST6 hindcast. We also conduct a background-sensitivity experiment using the substantially different WW3-ST4 hindcast forced by CFSR winds. As stated in the manuscript, the accurate but sparse altimeter observations and the dynamically evolving numerical background are both indispensable components of the fusion framework.

More fundamentally, the fusion method relies on a physically credible and skillful background field. In regions and periods where direct observations are sparse or distant, the fused estimate necessarily retains a larger contribution from the numerical hindcast. The accuracy and physical realism of the background therefore directly affect the quality of the fused product: a more accurate background will generally lead to a more accurate reconstruction, whereas a random, physically implausible, or severely biased field could not serve as a meaningful background. Our motivation is therefore to adopt a background field that is as "right" as possible, i.e., physically credible, skillful, and independently validated, rather than to assume that a "wrong" field can simply be repaired.

At the same time, no numerical wave simulation is expected to reproduce observations perfectly. Model-observation discrepancies may arise from multiple sources, including atmospheric forcing errors, source-term parameterizations, spatial resolution, numerical propagation, bathymetric representation, and omitted processes such as wave–current interactions. Determining the physical origin of each discrepancy and improving the corresponding model physics are important research objectives, but they are fundamentally different from reconstructing the most accurate historical SWH field possible. The present study addresses the latter objective.

The proposed product should also not be characterized as a simple statistical correction that merely masks model bias. Before fusion, we apply a one-dimensional look-up-table adjustment to remove the mean systematic difference between WW3 and altimeter SWH. The manuscript evaluates this step separately. It reduces the RMSE by only approximately 0.01 m and increases the correlation by only 0.001, demonstrating that simple bias adjustment contributes negligibly to the overall improvement. In contrast, the full space–time fusion reduces the RMSE against withheld Jason-2 observations from approximately 0.34 to 0.23 m and increases the correlation to 0.985. The improvement therefore cannot be explained as merely masking a mean model bias.

Furthermore, the fusion does not erase or conceal the original model information. The original WW3 hindcast and CCI Sea State observations remain independently available, while the fused field is provided as an additional derived product. Differences between the original hindcast and the observation-constrained reconstruction can themselves be analysed to identify where and when the numerical model systematically departs from observations. The dataset may therefore facilitate, rather than discourage, subsequent investigations of atmospheric forcing, source terms, and other model deficiencies.

The background-sensitivity experiment further demonstrates that the method is not an empirical correction tailored to a particular model configuration. WW3-ST4 and WW3-ST6 differ in both source-term parameterizations and atmospheric forcing. Although the fused result remains dependent on the background, as any model-assisted reconstruction must, the discrepancies between the two fused products are substantially smaller than the undersampling errors obtained when no physically meaningful background is used. This shows that the method requires a physically reasonable background but is not critically tied to one specific source-term package.

We respectfully disagree that the scientific value of a data paper depends on whether the dataset independently explains the physical origin of every corrected discrepancy. The purpose of the present dataset is to provide a spatially complete, observation-constrained reconstruction of historical SWH that mitigates both errors inherited from the numerical background and errors due to satellite undersampling. Such a product has direct value for wave-climate analysis, extreme-value assessment, model evaluation, training AI wave model, and applications that cannot be adequately supported by either sparse along-track observations or a standalone numerical hindcast. It is not presented as a replacement for physics-based model development, nor as evidence that further improvements to WW3 source terms are unnecessary.

Finally, we acknowledge that observation-model fusion and data assimilation only provides a pragmatic means of improving historical reconstructions and wave forecasting while numerical wave models remain imperfect. However, we see no basis for the suggestion that such products discourage improvements in model physics. On the contrary, the quality of the fused product depends positively on the quality of its numerical background, and improvements in source terms, atmospheric forcing, propagation, and other physical processes would directly improve future versions of the reconstruction.

Overall, we appreciate the reviewer's concerns regarding the physical consistency, potential applications, and limitations of the proposed dataset. However, several of the criticisms appear to arise from a different interpretation of the purpose and implementation of the study from that intended in the manuscript. The proposed product is neither generated by AI nor intended to replace physics-based wave modelling, or the original CCI Sea State observations. Rather, it is an offline, model-assisted and observation-constrained reconstruction of total SWH, designed to address the complementary limitations of sparse altimeter sampling and imperfect but spatiotemporally complete numerical hindcasts. The fusion method is explicitly formulated, its dependence on the numerical background is evaluated, and its performance is assessed against independent buoy and withheld satellite observations.

We have carefully reconsidered all of the reviewer's comments and have revised several expressions where the original wording could have been overly broad or insufficiently precise, particularly regarding the interpretation of temporal consistency and the intended use of the two-satellite product. At the same time, we believe that the main methodological and scientific concerns raised in the comments do not reveal fundamental flaws in the dataset, but largely reflect misunderstandings of its scope, methodology, and intended applications. We hope that the detailed responses above clarify these distinctions and provide a more accurate basis for evaluating the scientific contribution and appropriate use of the dataset.