Response Letter

Dear Editor,

Thank you for handling our manuscript submitted to GMD. We sincerely appreciate the time and effort that you and the reviewers have dedicated to evaluating our work. The reviewers' comments were highly constructive and have helped us improve the quality and clarity of the manuscript. In response, we have carefully revised the paper in accordance with their suggestions.

We hope that this version of the manuscript is acceptable for publication in GMD.

If you have any questions, please feel free to contact us.

We look forward to your positive response.

Sincerely,

Haoyu Jiang, Ph.D.
College of Life Science and Oceanography, Shenzhen University
Email: Haoyujiang@szu.edu.cn

<u>Response to Reviewer 3:</u>

For the record, this is the second version of the manuscript I have reviewed, I did not review the original version

I appreciate the effort of the authors to review revise the manuscript, and the structural and positive way they addressed my concerns. The remaining differences are more rooted in the fact that we are bringing two fields together here. Those of classical physical wave modeling and those of classical data driven modeling. This will lead to different angles of the problem we look at and sometimes different conclusions. In the bigger view of this, this is a discussion that will take time as the fields get to work better together. This paper should be seen as a leap forward in data driven wave modeling and should not be held back because as communities, we are learning each other's expertise and idiom. In that context these availability of our discussion in the GMD publication process should add to the core value of the publication by itself. With this I gladly recommend the present manuscripts for publication.

I have one minor suggestion to correct a correction I requested in the previous review and will expand on two of my previous comment considering the discussion provided by the authors.

We sincerely thank the reviewer for the careful reading of our revised manuscript and for the constructive and thoughtful feedback. We are very grateful for the recognition of our efforts in addressing previous concerns, and for the recommendation to accept the manuscript for publication. We especially appreciate the reviewer's insightful remarks on the broader context of classical physical modeling and data-driven approaches

1) WW3 references: I appreciate that the manual reference is replace by Tolman (1989). However, the latter is an internal report rather than a peer reviewed paper. A better foundational reference for the WAVEWATCH series of models would be the first peer-reviewed publication (Tolman 1991, JPO 21, 782-797). With WW3 using different governing equations a separate foundational reference would be appropriate, for which Tolman et al. 2002, Weather and Forecasting, 17, 311-333 is usually used.

We have revised the manuscript to replace the reference to the WW3 manual and the internal report (Tolman, 1989) with more appropriate peer-reviewed publications (Tolman, 1991; Tolman et al., 2002) We appreciate the reviewer's suggestion in helping us improve the rigor and completeness of our citations.

2) The nature of the ERA5 dataset. It is well-known that we have insufficient wave observations to create a reanalysis that is data-dominated. This implies that the ERA5 data is more accurate at locations where data is or has recently been available, and less accurate at locations where no recent data has been available. This results in an anisotropic error structure. The same can be expected to be the case for the data driven model presented here. If the data that has been assimilated in ERA5 is used to validate the data driven model you effectively validate the model where it inherently is most accurate. The resulting error will then not be representative for the entire domain. If, in contrast, the data driven model was to emulate a pure hindcast, it will likely show a larger error against independent data. That error, however, will be representative for the entire model. Whereas this subtlety may become important when the data driven models become mature, it is not a major issue for this initial (successful) attempt to create such a data driven model.

Regarding the nature of the ERA5 dataset, we are grateful for the reviewer's clarification of the error structure in reanalysis data and the implications it has for model validation. On this point, we may hold a slightly different view (though not necessarily the definitive one), which we offer for your consideration. The core principle of an AI model lies in learning statistical associations between inputs and outputs, rather than simply "memorizing" the input-output pairs. The non-uniform characteristics introduced by data assimilation—stemming from the randomness of both assimilation locations and observational errors—can be regarded as a form of random error. Consequently, training an AI model on data that contains randomly anisotropic errors does not inherently lead the model to learn or reproduce that anisotropy. In other words, a data-driven model will not inherit the anisotropic error structure present in datasets such as ERA5. Indeed, in our independent test set, the model has no means of knowing where data assimilation has occurred in ERA5, as it has never "seen" any ERA5 data from the test set.

Therefore, although ERA5 may exhibit spatiotemporal heterogeneity in error due to the sparse distribution of assimilated observations, such heterogeneity is unlikely to negatively affect the training of our AI model. On the contrary, data assimilation generally improves the overall accuracy of the model outputs, which is precisely why we chose ERA5 as the training dataset. In our previous revision, we also validated the AI model using independent buoy observations, which confirmed the overall good performance of of AI model.

3) My note on variance errors and a Taylor diagram. I agree that the variance error is defined by the error metrics used here. I mention it because representation of minima and maxima is a critical metric when a model is used by forecasters. The latter is closely related to the variance error of the model, and hence ahs value to be presented separately.

We also appreciate the reviewer's expansion on the importance of variance errors and the role of metrics such as those visualized in a Taylor diagram. We agree that accurately representing variability, especially extremes, is essential when models are applied in forecasting contexts. As mentioned in our previous response, while the Taylor diagram is indeed a valuable tool for model evaluation, scatter plots may provide a more detailed assessment when evaluating a single model. In particular, scatter plots can offer clearer insights into the errors associated with both minima and maxima, which are not explicitly captured by the Taylor diagram.

Once again, We sincerely thank the reviewer for the constructive suggestions and the positive recommendation on our revised manuscript.

Response to Reviewer 4:

I thank the authors for their detailed response and explanation. I mostly have minor editorial comments, which are made directly in the PDF of the manuscript (attached after this review report). Here, I summarize my main editorial comments (1-3) and list a final suggestion (4) regarding one easy analysis to strengthen the paper further. Once these comments are addressed, I think the paper will be ready for acceptance.

We sincerely thank the reviewer for taking the time to carefully read our revised manuscript and for providing such detailed and constructive feedback. We greatly appreciate your thoughtful comments, which have been extremely helpful in improving the quality of the paper.

A point-by-point response to the review comments is provided below. A substantial number of the suggestions made in the annotated PDF have also been incorporated into our revision. Please refer directly to the updated manuscript and the tracked-changes file for details.

We hope that our revisions have addressed all remaining concerns and that the manuscript is now suitable for acceptance.

1.Formatting:

-Citations in the Introduction: there are two types of format issues when citing references. a) LastName et al., YYYY should be used, but sometimes the comma after et al. is missed (e.g., on line 91, Ardhuin et al. 2019 ; on line 115, Hersbach et al. 2020). b) sometimes, "et al" is missed for references with multiple authors. For example, on line 144 and line 149, Liu (2021) is used, but should be Liu et al. (2021). I don't have time to mark out all the instances, but the authors need to check and fix thoroughly throughout the manuscript. It should be straightforward if the authors are using a citation tool.

-Change U10 and V10 to U_{10} and V_{10} ,which are more commonly seen in the literature.

We thank the reviewer for pointing out the formatting issues in the citations and variable notation. We have carefully gone through the entire manuscript and corrected all instances of improper citation formatting. We have also updated the notation of U10 and V10 to U_{10} and V_{10} in the manuscript.

2. Organize information presented in the Results section better.
- Breaking the Result section into subsections to help readers follow better: I added some suggestions directly on the PDF.
- For Fig. 2, I think it is easier to describe results with and without data assimilation in two consecutive paragraphs, without jumping back and forth between Fig. 2 and Fig. 3. So, I suggested moving the paragraph starting on line295 to after line 281.

Thank you for your advice, we have reorganized the content by breaking the results section into subsections to improve readability and flow. We have also followed your suggestion on the paragraph order: the discussion related to Fig. 2 with and without data assimilation has been restructured into two consecutive paragraphs. These changes help make the manuscript clearer, and we greatly appreciate your detailed feedback.

3. Reflect the response to some of my previous questions/comments in a concise sentence in the manuscript wherever the authors think applicable:
I appreciate that the authors did some extra testing to address my comments. While I agree that they do not need to include the figures attached in the response, it would be nice to add a concise sentence in the related section to show and assure readers that they have done robust testing. For example:

- Regarding my question about choosing 2020 as the test year, I suggest the authors adding "Note that evaluating against other untrained years yields similar results, with differences in correlation coefficient (CC) and root mean square error (RMSE) being less than 0.003 and 0.03, respectively.", perhaps at the end of Line 268.

- For other questions, I leave it to the authors to decide whether adding extra explanation to readers in one sentence would help improve the understanding/credibility of the paper.

We thank the reviewer for the thoughtful suggestion. Following your advice, we have added a concise sentence in the relevant section to indicate the robustness of our testing regarding the selection of the test year. For other related issues, we have carefully considered their relevance and clarity and, where appropriate, have provided brief explanations, as follows：

Added description in the subsection 2.2.3: "…Using our training samples (data from 2000 to 2017), training took approximately one hour per epoch on an NVIDIA RTX 4090 GPU. <u>It is also tested that the model's performance will be slightly worse if the training samples are significantly reduced.</u> Once trained, the model requires less than 10 minutes to compute (infer) the global SWH for one year at a spatio-temporal resolution of 0.5° × 0.5° × 1h on an NVIDIA RTX 3060 GPU."

Added description in the last paragraph of the subsection 3.2.1: "… A simple visual inspection of the movie indicates that the AI model effectively captures SWH evolution, suggesting that the AI model could serve as an effective surrogate for NWMs, at least for some wind-sea-dominated regions. <u>The spatial distribution of error metrics varies in different seasons, and such variability shows the same pattern in AI-based models and NWMs.</u>"

4. Regarding the attribution of the larger CC errors in the AI-model to swell pools: I think the authors can support this interpretation further (beyond citing references) by showing (or

binning) the error metrics as a function of the misalignment angle between wind and dominant or mean waves. Following the authors' argument, we could expect that on average, the CC error metric reduce with this misalignment angle between wind and waves (absolute value from 0° to 180°). Note that 0-45 is aligned/wind seas, 45-135 is cross-seas, and >135 is swell opposing wind at least commonly used in the hurricane-wave community (Holthuijsen et al. 2012; https://doi.org/10.1029/2012JC007983). I think this should be easy to do, since the mean wave direction can be obtained from ERA5 (and likely also from CCI-Sea State dataset).

I think adding this information in a figure will strengthen this paper. In fact, one concern I still have with the current version of the manuscript is that the results section is populated by a bunch of figures in the same manner simply for different data sources or with/without data assimilation (Fig. 4-9). (It is a bit excessive such that it seems Fig 7 and 8 can also be supplementary information.) If the key message of Fig 7 and Fig 8 is just that data assimilation reduces the overall errors, especially in the swell-dominated regions. It could probably also be summarized by the extra figure I suggested above (i.e., plotting the error metrics as a function of the misalignment angle between wind and dominant or mean waves). The authors would also need to keep in mind my comment #2 above when adding this figure. The order of information in the Results section may need to be slightly re-arranged if this extra figure is included.

We appreciate the reviewer's insightful suggestion regarding the attribution of CC errors to swell-dominated conditions. Following your recommendation, we added an analysis showing the variation of correlation coefficient (CC) as a function of the swell energy proportion, which we believe provides a more direct and reliable indication of swell influence than the misalignment angle between wind and wave directions due to several reasons:

1) In many ocean regions, multiple wave systems (or partitions) often coexist. In such cases, neither the mean wave direction nor the peak wave direction can fully represent the overall wave field, and relying solely on these metrics may even be misleading.

2) Even when wave energy is concentrated in a single direction, the alignment between wind and wave directions does not necessarily indicate that the sea state is wind-sea dominated. The classification of wind sea versus swell typically requires an analysis of the relationship between wind speed (sometimes, its component projected onto the wave direction) and the phase speed of the waves.

3）When wind and wave directions are misaligned, a directional difference of 180° is not inherently more indicative of swell conditions than a difference of 90°.

ERA5 provides separate records of wind sea and swell significant wave heights based on spectral partition technology, allowing us to compute the swell energy proportion as $SWH_{swell}^2/SWH_{total}^2$. We have added a new figure to illustrate how the model performance (in terms of CC) changes with this ratio. The new figure and corresponding explanation have been added to the corresponding paragraphs, the revised manuscript is described as follows:

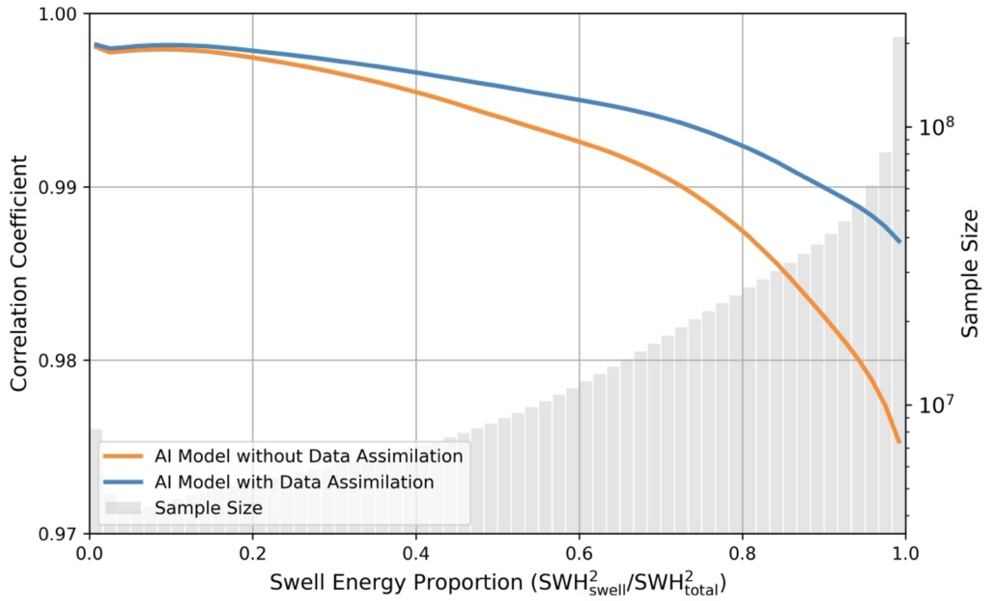

Figure 5. Correlation coefficient between AI model and ERA5 data as a function of swell energy proportion the global ocean in the year of 2020. The orange and blue lines represent the AI model before and after data assimilation, respectively, and the grey bars indicate the variation in sample size as a function of swell energy proportion.

"…However, in the tropical oceans, especially along their eastern coasts where swells are predominant ("swell pools", Chen et al., 2002), the CCs are below 0.9 (~0.85 in the Indian Ocean, ~0.8 in the Atlantic Ocean, and ~0.7 in the Pacific Ocean). To further examine whether these results are related to the presence of swell, we examined the relationship between the swell energy proportion (i.e., the ratio of the square of the swell SWH to the square of the total SWH) and the CC across the global ocean (the orange line in Fig. 5). The results show a clear trend: the smaller the swell energy proportion, the higher the CC. In particular, when the proportion is below 0.7, the CC values are consistently above 0.99, indicating robust model performance in wind-sea-dominated regions. However, when the swell energy proportion exceeds 0.7, the CC values for the model without data assimilation drop significantly, corroborating its lower performance in swell-dominated regions."

And added description in the last paragraph of the subsection 3.2.2: "…the overall errors become significantly smaller, particularly in the swell-dominated regions, with assimilation. These results are further supported by Fig. 5, where the model with data assimilation consistently maintains substantially higher CC values, even when the swell energy proportion exceeds 0.7. Similar to Supplementary Movie S1, the comparison animation of the results with assimilation is placed in Supplementary Movie S2, …"

And as per your recommendation, we have transferred the original Figs. 7 and 8 to the Supporting Information, where they are now designated as Figs. S7 and S8.

This comment is from annotated PDF (Line 285): ***the same should also be true for the hot-start experiment, right? So it is interesting that this variability in the error metrics is also larger in the cold-start experiment before errors stabilize (blue shading being wider than the orange).***

Yes. As you point out, the variance of the error metrics is indeed wider in cold-start experiments, especially in the early stages of prediction before the error stabilizes. This wider spread (blue shading) reflects the fact that cold-start methods are more sensitive to the initial wind field due to the lack of a priori information. In contrast, hot-start experiments benefit from the availability of a wave initial field, which can produce more stable results.