Response Letter

Dear Editor,

Thank you for concerning our manuscript in GMD. We appreciate you and the reviewers for your earnest work. The comments from the reviewers are very helpful, and the paper has been revised carefully according to these comments. Our point-by-point responses to the comments of the reviewers are attached, as well as a tracked-changes version of the manuscript. For the comments that we do not completely agree on (only a few), we also give our explanations in the point-by-point responses.

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 appreciate your support very much.

Thank you for your time and consideration. 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

Response to Reviewer 3:

For the record, this manuscript was sent to me for review as a revision. This is the first time I have seen and reviewed the manuscript.

The authors present a data driven global wave model where a present wave height and a wind speed 6h ahead in time are used to estimate wave height 6h in the future. A "rolling model' is created by repeating this step, and data assimilation is added to make the model more accurate. The only thing in these sections that should be highlighted more is how the data assimilation is performed. On line 185 it is stated that Data Assimilation is used to "correct the model's "initial" SWH field.", whereas later (line 366) it is stated that data is assimilated "every 6 hours". Before going into more detailed critique, it needs to be stated that resulting model and its analysis are clearly suitable for publication in GMD.

Dear Reviewer:

We would like to thank you for dedicating time to carefully read our manuscript and provide feedback. We sincerely think your detailed comments have helped us to improve the manuscript, and revisions are made according to them. A revised version of the manuscript with changes highlighted is also attached to this response letter. We hope that the revised version of the manuscript meets your expectations. For the few comments where we hold a different perspective, we have provided detailed explanations in the following point-by-point responses.

We would like to clarify that our model is a rolling forecast model with the simplest 1-hour-by-1-hour time steps. Specifically, the model takes the SWH field at time T_i and the wind field at T_{i+1} as inputs to predict the SWH at T_{i+1} , which is similar to numerical wave models (NWMs). We believe this has been clarified in Section 2.2.1. In this setting, the outputs of the last time step will be the "initial" field of the next time step, and the frequency of data assimilation can

be user-defined. Here, assimilation was conducted every 6 hour in our experiment, but a higher or lower assimilation frequency can also be used. Generally, a higher assimilation frequency leads to more accurate results but also entails increased computational costs, and vice versa.

To better clarify this, we added some explanation to the revised manuscript:

"…we tried to incorporate data assimilation techniques by integrating altimeter measurements to correct the model's "initial" SWH field. It is noted that in our input-output setting, the outputs of the last time step will be the "initial" SWH field of the next time step…."

"In our data assimilation experiment, assimilation was conducted every six hours (i.e., every six time steps, observations are used to corrected the outputs of the rolling model and the updated outputs are used as the new inputs at the next time step), beginning after the first 24 hours of the model run. Of course, the frequency of data assimilation can be user-defined. A higher assimilation frequency generally leads to more accurate results but also entails increased computational costs, and vice versa."

The input and target of the study is the ERA5 reanalysis. Note that this reanalysis consists of a wave hindcast with most altimeter data assimilated into it. Note that due to the lack of sufficient wave data to generate a data dominated wave analysis, ERA5 still is mostly a wave model hindcast, and that its quality is inhomogeneous. As ERA5 is more accurate at the locations of the assimilated altimeter data, validating with ERA5 and cross referencing with the same altimeter data is a little incestuous and produces error measures that are too rosy. For this reason, I would have preferred developing the AI model with hindcast without DA, and then comparing the AI model, the input model and the altimeter data in a three-way approach would be a cleaner analysis of various data sources. This is not disqualifying for the present study, but the limitation (validation with dependent data) needs to be discussed.

We acknowledge the problems of ERA5 that you mentioned. In spite of these problems of ERA5, training an AI model requires high-quality input data to achieve reliable results. From this perspective, ERA5 is still a good dataset to train against.

Regarding the independence of result evaluation, we respectfully disagree the comment that "validating with ERA5 and cross referencing with the same altimeter data is a little incestuous and produces error measures that are too rosy". If the comparison is made between ERA5 and the altimeter data that has been assimilated to ERA5, it will be, of course, incestuous or even unreasonable. However, the comparison is made between the AI model and ERA5, and between the AI model and altimeter measurements. We need to emphasize that once the training of the AI model is finished, it can be regarded as a model logically independent of ERA5 (or any hindcast dataset it is trained against). We can draw an analogy between the training process of AI models and the tuning process of NWMs, as both are essentially adjusting a set of empirical coefficients (though AI models typically have far more parameters to "train"). Specifically, we can tune the NWMs using ERA5 reanalysis data or directly with altimeter observations, and then validate the NWM results against ERA5 reanalysis or altimeter observations in a different periods (e.g., altimeter measurements from years not used in tuning). Here, the training and validation process of our AI model strictly follows this same logic. Therefore, we believe it will not introduce problems when using the ERA5 data as the training target and testing benchmark (of course, the training set and testing set should be separated), and the comparison made in this study is totally reasonable and will not generate rosy error metrics.

Certainly, we by no means suggest that ERA5 constitutes the optimal training dataset for developing such AI models. On one hand, there undoubtedly exist better methodologies for calibrating NWM hindcast or assimilating/merging observations into NWM hindcasts. On the other hand, NWMs themselves are continually evolving, with their output data achieving progressively higher accuracy. The more fundamental objective of this manuscript remains

investigating the feasibility of this "simplest" input-output strategy in AI modeling. The quality of ERA5, according to our results, is adequate for this purpose.

We hope this explanation meets with your approval. Should you hold differing views, we warmly welcome your further comments in the review report and would be pleased to continue the discussion on this matter.

As a traditional wave modeler who has also worked in AI for decades, I find the justification for doing an AI model weak. On line 9 it is stated that wave models "are computationally intensive and constrained by incomplete physical representations of wave spectral evolution.", yet the ERA5 data used here is founded in these limited models. Moreover, wave models provide much more data than the SWH for practical wave predictions, or in coupled environmental models. Yes, NWM are more expensive, but we have been providing operational forecasts since the 1960s with such models (contrary to suggestions provided on line 36), so apparently the expenses are not prohibitive. The paper will be stronger without half-baked justifications.

We completely agree with the reviewer that NWM also has many (more) advantages compared to contemporary AI wave models (including the one in this study). Here, we list the limitations of NWM merely to demonstrate that the proposed AI model still has its merits that can overcome some of the NWM's problems, not to claim that the AI model outperforms NWM, let alone suggest it could replace NWM.

Although the ecWAM which ERA5 is based on also has the problem of incomplete physical representations of wave spectral evolution and numerical effects (e.g., discrete interaction approximation and garden sprinkler effect), these effects can be alleviated if enough data is assimilated/merged to the model output and assimilation can generate a more reliable analysis field. This is why today's AI weather forecasts nowadays can beat numerical ones in some error

metrics by training against ERA5. Using the data combining observations and NWM outputs, we also realized a AI model that has better accuracy than NWMs with respect to SWHs recently (with a different input-output strategy) (Wang and Jiang 2024). However, we have to admit that the aim of this manuscript is not to overcome the incomplete physical representations and numerical effects, so we just simply mention this problem in the introduction.

To provide a balanced perspective on both NWMs and AI-based wave models, we have added clarifying statements in the introduction's concluding section to address your concerns:

"Although good results have been obtained by the AI model presented in this study, it is noted that we do not intend to suggest that the AI model is superior to traditional NWMs or that it could replace NWMs. NWMs still retain numerous advantages over AI approaches, such as their ability to provide parameters beyond SWH and their stronger physical interpretability, among other merits. The AI model we have developed should be more regarded as a model surrogate specifically for time- or computation-sensitive scenarios."

Ref.:

Wang, X., & Jiang, H. (2024). Physics-guided deep learning for skillful wind-wave modeling. Science Advances, 10(49), eadr3559.

Considering that wind waves dominated the higher wave heights and the overall errors worldwide, a model like this AI model that is focused on representing wind seas should result in reasonable wave heights but also will have issues in areas with multiple (dominant) swell fields. This is acknowledged in the manuscript in discussing the errors in "swell pools". It would be nice to acknowledge the need of being able to do swell accurately too for many applications. Since the "waves across the Pacific" studies in the 1960s, it is well known that NWMs can do this well. Note that with the dominance of wind seas in model errors, the differences between cold and hot started results, as well as impacts of DA are at least

qualitatively expected. Note that this could be expected as present wave height and future wind speed allow for assessing where the dominant wind wave field is with respect to growth stage of the wave field. Did you check if additional accuracy can be attained if the present wind field is used too (this would be a proxy for wind sea and swell separation in the initial wave height in the algorithm)?

As noted, the model performs well in wind-sea conditions but less so in swell regions, as discussed in the "swell pools" error analysis. Undeniably, swells play a crucial role in many applications, thus, there are needs of being able to model swell accurately. However, although the propagation of swells has been well understood since the famous field experiment, 'Waves Across the Pacific', they their behavior retains characteristics of an initial-value problem, making swells a persistent source of error in current NWMs (e.g., Jiang et al. 2016). Therefore, NWM also usually perform worse in swell-dominated regions than in wind-sea-dominated regions. From Figure S4 in the Supporting Information, it can be seen that NWM demonstrates similarly low correlation coefficients in "swell pool" regions, which is only marginally higher than our AI model (as shown in Figure R1 below). This explains why we consider the AI model's performance in swell-dominated regions is still acceptable.

Regarding the suggestion to incorporate the present wind field as an additional input, we made some tests according to the reviewers suggestion. However, we found that it did not lead to improvement compared to the current model. This is not surprising because the variation of wind is usually very small within one hour so that the pattern of the present wind field and the 1-h future wind field is very similar. Moreover, although wind-sea and swell SWH can be roughly separated using wind speed + SWH information (which we believe is exactly the reason for our AI model can still model swells with acceptable accuracy), the evolution of swell SWH is also dependent on the direction and period of swells.

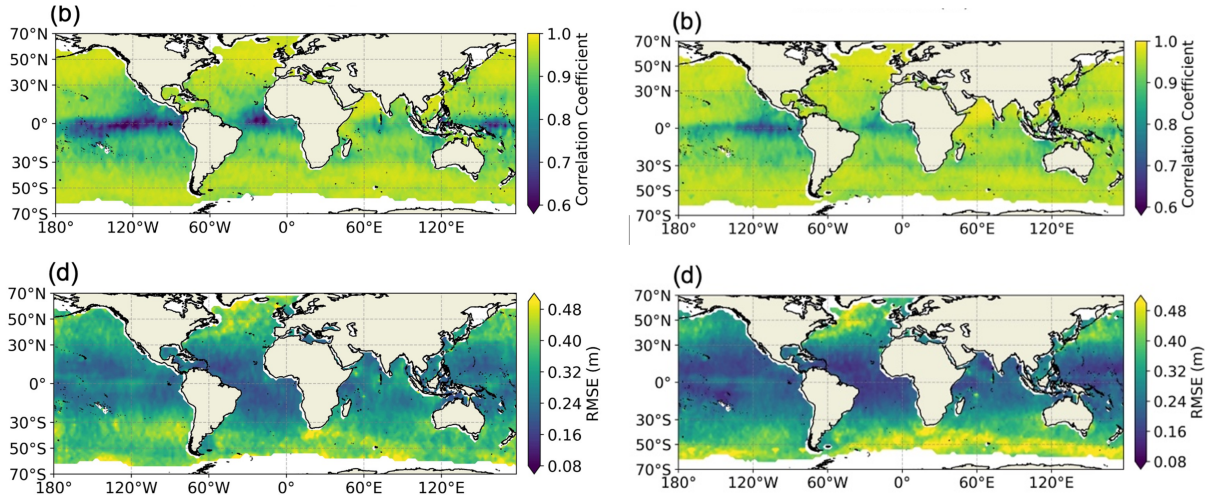

Figure R1. *The spatial distributions of correlation coefficients (upper) and RMSE (lower) between model results and CCI-sea state data in 2020 for global ocean: (left) AI model V.S. CCI-sea state, and (right) WW3-ST6 NWM V.S. CCI-sea state. The left column is from Figure 5 in the manuscript and the right column is from Figure S4 in the Supporting Information.*

While adding the present wind field is insufficient for swell modelling, adding additional historical wind fields can indeed improve swell modelling, as demonstrated in our previous paper (Wang and Jiang 2024). This enhancement stems from the nonlinear teleconnection between swell energy and distant historical wind forcing. However, with increased wind field inputs, the significance of the initial SWH field diminishes substantially, while the model's physical framework and input-output relationships undergo fundamental modifications. Consequently, in the present study, we maintain our methodological focus on the simplest rolling modelling approach, i.e., one wind field + one SWH field.

Ref.:

Jiang, H., Babanin, A. V., & Chen, G. (2016). Event-based validation of swell arrival time. Journal of Physical Oceanography, 46(12), 3563-3569.

Wang, X., & Jiang, H. (2024). Physics-guided deep learning for skillful wind-wave modeling. Science Advances, 10(49), eadr3559.

Specific comments:

Line 17: "… the errors of the model diverge lightly …". Perhaps use "accumulate" or "increase" as this is not divergence in the classical meaning in environmental sciences.

Thank you to the reviewer for pointing this out and we have changed the words from "diverge/divergence" to "accumulate/accumulation" in the revised manuscript.

Line 20: "This deep learning model can not only serve as an efficient surrogate for traditional numerical wave models but also provide a baseline for statistical modeling of global SWH due to its simplicity in inputs and outputs." For model uncertainty, where generally only wave heights are considered, this indeed could be a good application.

We appreciate your positive recognition of the potential application of our deep learning model in providing a baseline for statistical modeling of global SWH. To better stress this model is only for wave height, this sentence is slightly revised to :

"This deep learning model can not only serve as an efficient surrogate for traditional numerical wave models with respect to SWH but also…"

Line 32-33: WW3 is referred to by its manual, SWAN by foundational papers. Please balance your references (I would prefer foundational references, and manuals only when the model is used to identify the version).

Thank you for pointing this out and we have changed the citation：

Tolman, H. L.: The numerical model WAVEWATCH: a third generation model for hindcasting of wind waves on tides in shelf seas, Delft University of Technology, Department of Civil Engineering, Fluid Mechanics Group, Delft, 1989.

Line 114: The WW3 data is not used here at all. Why is it in the materials section then?

We did compare the performance of the AI model with the WW3-ST6 in the part of discussing the error of 'swell pools'. This is why we can say that our model is comparable to those of state-of-the-art NWMs, and why we can say AI model performs well across global oceans in general, both in wind-sea- and swell-dominated regions (although the expression "both in…" is removed in the revised manuscript). To save the number of figures in the text and to make the manuscript more reader-friendly, we put the figures of the results of the comparison between the WW3-ST6 and the CCI-Sea State in the Supporting Information Figure S4.

Line 185: DA use for "initial" wave height is misleading, as data is assimilated every 6h (Line 366).

Thank you for pointing this out. To better explain this, we added a sentence after this part: "It is noted that in our input-output setting, the outputs of the last time step will be the "initial" SWH field of the next time step." This should eliminate the potential misunderstandings.

Line 193: Adding a measure for the representation of the signal such as variance of the wave height (as in a Taylor diagram) would be useful, since minimizing a rms error tends to result in a smooth model that smooths out some highs and lows.

We are a bit confused about this comment as Line 193 (equations) seems to be not relevant to your comment. We assume you are talking about the error metrics. However, when the RMSE and correlation coefficient are known, the location in the Taylor diagram is determined.

Therefore, we think the four error metrics used in this study is sufficient to describe the error property. In our opinion, Taylor diagram is more suited for the comparison of several different models, but here we only have AI model and WW3-ST6, thus, we do not feel the necessity of using a Taylor diagram.

Also, minimizing the RMSE does not necessarily tend to result in a model that smooth out highs and lows. This only happens only when the problem is too complex for the model to accurately capture the high and low variations. When it comes to the modelling of SWH, we believe this poses even less of a concern because the variation of SWH is usually smooth itself since SWH can be regarded as a "low-pass filter" of winds.

If we misunderstood this comment, it will be nice if you can expand it a bit.

Figure 2: It appears from the wavy behavior in the DA runs that data is created every 3 hours, not every 6 hours are claimed in the description of the AI model. This needs to be explained before publication.

We confirm that data assimilation is performed every 6 hours, as described in the manuscript, and this can also be verified in Figure 2. For example, in Figure 2(a), the red box highlights a 24-hour period. Within this period, the blue curve exhibits 4 distinct upward steps, each corresponding to an assimilation event. These steps indicate the corrections applied to the model every 6 hours. We hope this clarification addresses any concerns about the assimilation frequency.

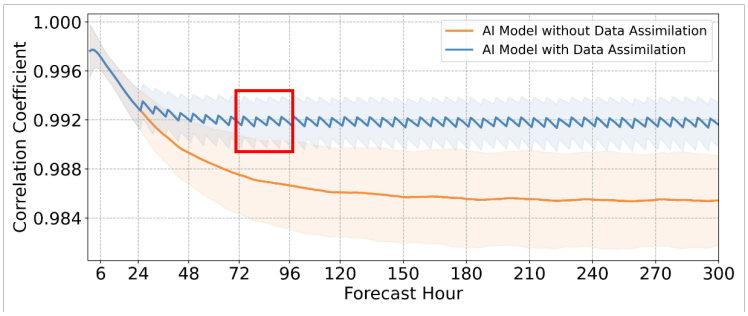

Figure R2. *Figure 2a in the manuscript. To better show the wavy behaviour of the curve, we use a red rectangular to show the four upward steps within 24 hours. The lines represent the mean values of the error metrics for the experiments starting from different initial SWH fields. The shaded areas around the lines indicate the range of error metrics across different experiments with varying initial SWH fields.*

Line 250: Note that the general statement about the effects of DA is accurate, but does not acknowledge that validating with the same but sparse data is not likely to be representative for areas without a recent observation.

We are also a bit confused about this comment. Are you suggesting that sparse altimeter observations are insufficient for evaluating areas without a recent observation? However, it should be emphasized that these results derive from 236 parallel experiments, in which the altimeter data can be considered effectively global in coverage. Moreover, most altimeters complete an Earth orbit in under two hours, resulting in spatial dislocations between the validation dataset and previously assimilated altimeter observations. To further address this issue, we included independent in-situ buoy observations for further validation (Figure 6 and S7 in the revised manuscript). The results of the comparison are in good agreement with those obtained from the altimeter data, supporting the robustness of our model.

Again, if we misunderstood this comment, it will be nice if you can expand it a bit.

Line 294: many years of experience with the GSE has taught us that the GSE results in unrealistic wave fields but has little impact on error statistics. I find this argument weak for that reason.

We respectfully maintain a differing perspective: when wave fields appear unrealistic, this indicates that data errors/inconsistencies have become substantial enough to be visually identifiable. In such cases, maybe one should no claim that "the error statistics have little impact on error statistics". In many scenarios, since SWHs in swell-dominated regions are typically smaller, the absolute errors (e.g., RMSE) induced by the GSE may appear small. However, this often corresponds to significantly increased relative errors, manifested through decreased correlation coefficients. While the random errors from GSE may not be as consequential as the first two factors mentioned, we contend they nevertheless make non-negligible contributions to the overall error.

Lines 326-333: The statements on the first and last lines about swell and wind seas do not seem to be consistent.

These two statements are not contradictory. The first statement conveys that the AI model demonstrates reasonably robust performance across both wind-sea and swell conditions. The last statement indicates that even if readers consider the swell performance less optimal, they should at least acknowledge the model's capability in wind-sea scenarios—where it could serve as an effective surrogate for NWMs.

To clarify this point and prevent potential misinterpretation, we have further removed the phrase "both in wind-sea- and swell-dominated regions."

Response to Reviewer 4:

Summary:

This study constructed an AI model for predicting the significant wave height (SWH) parameter globally using a convolution neural network with the U-Net architecture. The AI SWH model is trained on 18 years of ERA5 reanalysis by using the SWH and the 10-m surface wind vector fields at two consecu6ve 6mes (i.e., rolling prediction strategy). Therefore, the AI model "simulates" SWH in a manner similar to the numerical wave models with an initial SWH field and the forecasted 10-m wind fields. Evaluation of AI SWH model performance in 2020 shows that this AI SWH model performs as good as the WaveWatch III model with the ST6 physics. The global error patterns against ERA5 SWH and CCI-Sea State analysis product further show that the AI-SWH model produces more reliable SWH prediction in wind-sea conditions than in swell-dominant conditions. The authors conclude that this AI SWH model can be a more efficient approach to produce global forecast of significant wave height than traditional numerical wave models.

Dear Reviewer:

We would like to thank you for your patience in reading the paper in detail and your valuable comments. We sincerely think your detailed comments have helped us to improve the manuscript. Below, we present our point-by-point response (text in black denotes our replies). We hope the manuscript is now acceptable following our revisions and explanations.

Major comments:

Introduction:

My impression is that the introduction somewhat overstated the powerfulness of AI model or AI SWH model. It is true that the numerical wave models have limitations in

parameterizations of the wind input term and the dissipation term that govern the spectral evolutions. But I don't think the AI model are completely free from these limitations since it learns from ERA5 and inherently adopts those limitations the authors stated. I suggest the authors toning down a bit this aspect when writing about the advantages of the AI model and not giving an impression that the AI model alone could overcome the physical limitations of the numerical wave models.

We completely agree with the reviewer that numerical wave models (NWMs) also has many (more) advantages compared to contemporary AI wave models (including the one in this study). Here, we list the limitations of NWM merely to demonstrate that the proposed AI model still has its merits that can overcome some of the NWM's problems/limitations, not to claim that the AI model outperforms NWM, let alone suggest it could replace NWM.

To provide a balanced perspective on both NWMs and AI-based wave models, we have added clarifying statements in the introduction's concluding section to address your concerns:

"Although good results have been obtained by the AI model presented in this study, it is noted that we do not intend to suggest that the AI model is superior to traditional NWMs or that it could replace NWMs. NWMs still retain numerous advantages over AI approaches, such as their ability to provide parameters beyond SWH and their stronger physical interpretability, among other merits. The AI model we have developed should be more regarded as a model surrogate specifically for time- or computation-sensitive scenarios."

Thinking about the results from a more physical perspective:
It is quite interesting that the AI model is skilful in predicting the SWH associated with wind seas. I am just curious if this means that the AI SWH model has learned some physics of the wave evolution. Could the authors comment on whether this AI model be run in an idealized

setup to produce the SWH of fetch-dependent wind waves under constant and uniform wind forcings at different wind speeds? Would the relationship between SWH and U10 in this AI model (i.e., (SWH-U10)AI) behave similar to some empirical relations between U10, fetch, and SWH? For example, for fully developed seas, I think the authors can compute the SWH associated with the Pierson-Moskowitz spectrum at different wind speeds and obtain a SWH-U10 relationship predicted by the Pierson-Moskowitz spectrum (i.e., (SWH-U10)PM). For fetch dependent seas similarly, (SWH-U10)JONSWAP can be found for different fetches.

Thank you for this insightful comment.

The AI model is trained to learn the statistical relationships present in the training dataset rather than explicitly solving physical equations governing wave evolution. From this perspective, one can say that the AI model has learned some physics of the wave evolution, from a statistical point of view. In particular, our cold-start experiments demonstrate that when driven by realistic wind fields, the AI model progressively produces results closer to ground truth. This suggests, to some extent, that the AI has learned quantitative patterns of wave growth (labelling these as physics may be inappropriate—what the AI discovers remains fundamentally statistical in nature). Furthermore, SWH data in current wave models/reanalyses fundamentally adhere to the statistical relationships between U10 and SWH. When the AI model produces results consistent with these wave models/reanalyses, it implicitly indicates that these established relationships are largely preserved within the AI model.

However, we need to note that the AI model works in a different way from NWMs. After the AI model finish its training, the input form of the AI model needs to be exactly the same as that used in the training. In our case, the input of the our AI model has to be global SWH at T_i and global wind field at T_{i+1}. Some important but constant information, such as bathymetry and coastal morphology, and even the curvature of the Earth, are implicitly embedded in the AI model in a statistical way. Therefore, from our understanding, it seems to be impossible to

conduct idealized tests that can be easily done by NWMs, such as fetch-limited and duration-limited tests, in our AI model framework. This is also a limitation of the AI model. In our AI model, given the global domain of simulation, even prescribing a spatially "uniform" wind direction in the input fields is inherently unfeasible due to the spherical effect. Similarly, the AI model is not suitable for certain toy model experiments, e.g., we cannot setup a simulation of global SWH on an Earth without any land, which NWMs can easily handle.

Specific comments:

Methods:

1. How long does it take to train this AI Model on 18 years of data? Would it be fair to mention this training time as well?

We have mentioned the training time in Section 2.2.3 (Model Training) in the revised manuscript, which reads: "We used six batches for training and trained the model for up to 30 epochs at a learning rate of 0.0001 using the AdamW optimizer. To alleviate overfitting, we implemented a commonly used deep learning technique where training is halted when the loss in the validation set does not decrease for four epochs. Using our training samples (data from 2000 to 2017), training took approximately one hour per epoch on an NVIDIA RTX 4090 GPU. " Therefore, it takes less than ~30 hours to train this AI model using the data from 2000 to 2017.

2. Would the results changes if testing was conducted using data from 2018, 2020, and 2021 together? Have the authors tested how sensitive this model is to different ratios of the training data, the evaluation data, and the model testing data? Can authors provide some answers to these questions in the method?

According to the your suggestion, we also conducted the model test using the data from 2018, 2019, 2020, 2021. The following Figure R3 shows the presents the error curves of the AI model

on the test sets from 2018, 2019, 2020, and 2021. From this figure, it is evident that the model performs consistently across these three test sets, with differences in correlation coefficient (CC) and root mean square error (RMSE) being less than 0.003 and 0.03, respectively. Such differences are similar to the difference of error metrics for NWMs across different years. This results demonstrate that the model exhibits strong robustness and generalization ability across different test periods. This results align with our expectations: a properly developed statistical model, when evaluated on unseen test data, should demonstrate consistent performance across different years.

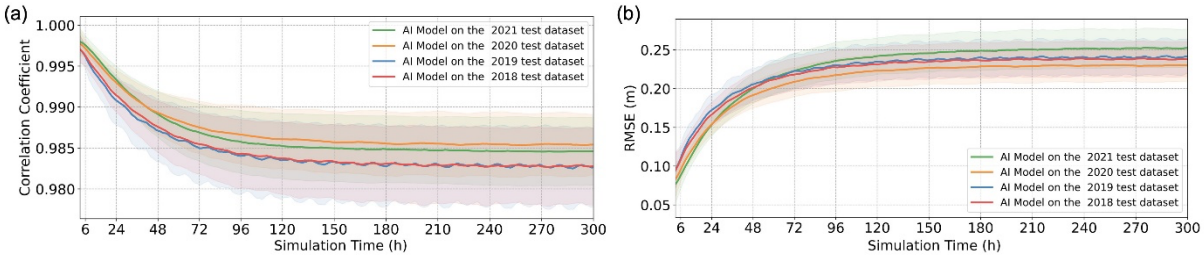

Figure R3. *The variation of global overall error metrics between the AI SWH model (training with the data of years 2000-2017) outputs and ERA5 with simulation time using data of different years as the testing set: (a) CC, (b) RMSE. The lines represent the mean values of the error metrics for the experiments starting from different initial SWH fields. The shaded areas around the lines indicate the range of error metrics across different experiments with varying initial SWH fields.*

Similarly, following your comments, we tested how sensitive this model is to different amount of the training data. According to the basic knowledge of deep learning, the ratio among the three datasets is not critical. However, insufficient training data volume may indeed lead to either overfitting or underfitting issues. Figure R4 shows the error curves of the AI model trained with different amounts of training data and evaluated on the 2020 test set. It can be seen that the model performance increase with the increase of the size of the training dataset.

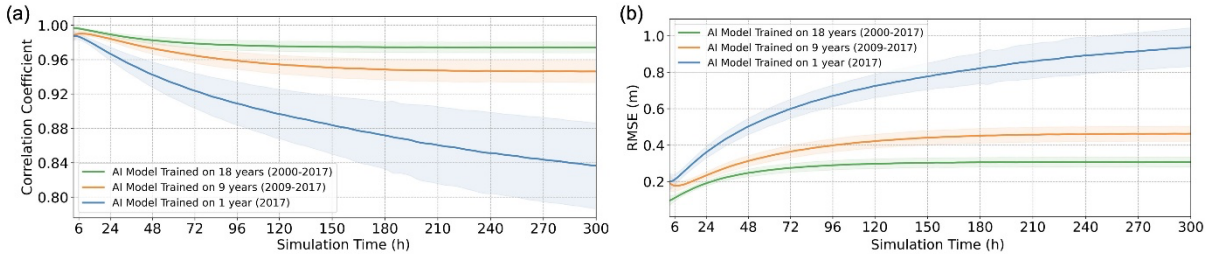

Figure R4. *The variation of global overall error metrics between the AI SWH model outputs and ERA5 with simulation time using data of different periods as the training set (testing using the data of year 2020): (a) CC, (b) RMSE. The lines represent the mean values of the error metrics for the experiments starting from different initial SWH fields. The shaded areas around the lines indicate the range of error metrics across different experiments with varying initial SWH fields.*

The results presented in both Figure R3 and Figure R4 are within expectations. We have not included the these findings and their associated discussion in the revised manuscript because we intentionally avoid delving into how technical details of the AI models (e.g., training data volume, model architecture, hyperparameters, number of layers and parameters) influence the model results. To some extent, when given the predefined input-output framework, there will almost always be opportunities - however marginal - for performance improvement of AI models through such technical refinements. We are not saying these technical details lack importance, but these details primarily represent engineering challenges in model implementation: through extensive experimentation, one could systematically explore which specific model architectures, hyperparameters, and training datasets might yield optimal results within this input-output framework.

3. Did the authors perform some model tuning based on the evaluation dataset? if so, it would be great if the authors document what parameters have been tuned using the validation set from 2022.

No, we did not perform any model parameter tuning based on the validation/testing set.

4. Also, it is not very obvious to me how or why choosing 2022 for validation can prevent overfitting. Could the authors demonstrate that this AI model is not overfitting in some way?

We employed the commonly used early stopping strategy in deep learning to alleviate overfitting. Specifically, during training, we monitored the mean squared error (MSE) on the validation set, and if the MSE remained unchanged or started to increase for several consecutive epochs, training was terminated to alleviate overfitting to the training data. We realized that we forgot to mention this in the manuscript, which is our problem, this has been added to manuscript:

"We used six batches for training and trained the model for up to 30 epochs at a learning rate of 0.0001 using the AdamW optimizer. To alleviate overfitting, we implemented a commonly used deep learning technique where training is halted when the loss in the validation set does not decrease for four epochs."

In principle, the choice of a validation set is flexible as long as it does not overlap with the training or test sets. Our decision to use 2022 as the validation set was based on two key considerations: (1) The year 2022 is temporally distant from the test set, allowing for a more objective assessment of the model's generalization ability and helping to reveal potential overfitting issues. (2) There is no corresponding CCI altimeter data for 2022, meaning that this year's data could not be used for other parts of our analysis (comparing with CCI data). This made it a natural choice as an independent validation set.

Even with the implementation of early stopping strategy, we cannot conclusively demonstrate that the AI model is not overfitting. Your comments rightly reminded us that replacing "prevent" with "alleviate" would constitute a more precise formulation.

Results:

1. Figure 2: With data assimilation, why do the time series of the 4 error metrics have a zigzag pattern?

This pattern is simply a natural consequence of the data assimilation process, which is performed every 6 hours, using altimeter observations to correct the SWH output of the last time step and using the corrected SWHs as the input of the next time step. After each assimilation, the accuracy of the model output improves as errors are corrected. Then as rolling model continues, small errors continue to accumulate, leading to a gradual decline in accuracy until the next assimilation step.

2. Figure 4: Do the spatial distributions of the 4 error metrics change in different seasons?

Sure, the spatial distributions of error metrics will change in different seasons, because the wave climates are different for different seasons. Such a change can be observed in all wave models, such as simple statistical models, AI models, and NWMs.

Figure R5 illustrates the error curves of the AI model in different seasons during the rolling inference process. The results clearly indicate seasonal differences in errors: the overall errors are lowest in JJA and highest in DJF, although the difference is generally small.

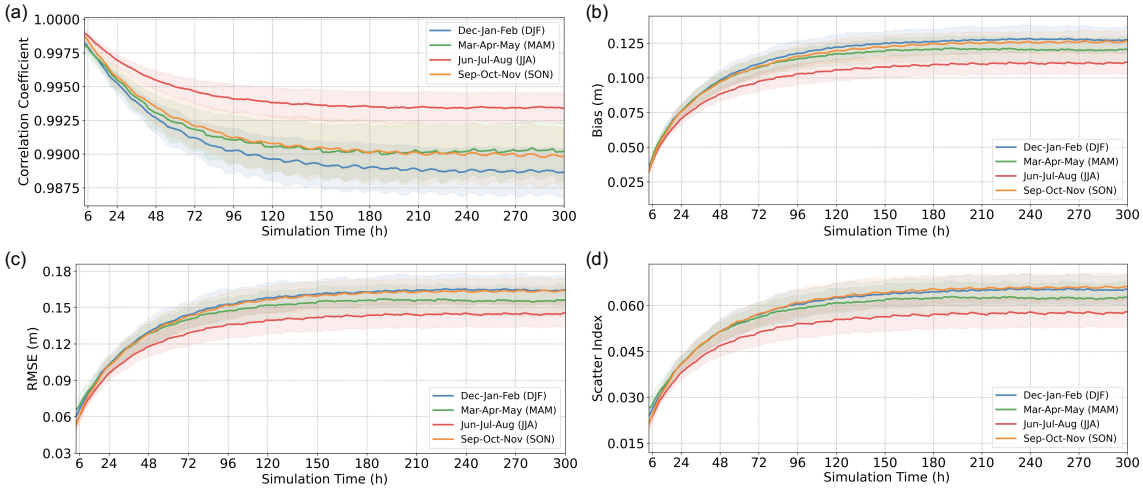

Figure R5. *The variation of global overall error metrics between the AI SWH model outputs and ERA5 with simulation time in different seasons: (a) CC, (b) bias, (c) RMSE, and (d) SI. The lines represent the mean values of the error metrics for the experiments starting from different initial SWH fields. The shaded areas around the lines indicate the range of error metrics across different experiments with varying initial SWH fields.*

We also examined the spatial distribution of the AI model's 240-hour hindcast errors, as shown in Figure R6. The error patterns are also different in different seasons, which is linked to the strong seasonal variations in wave climate. However, these patterns all show that the model perform well in wind-sea dominated regions and the performance degrades in swell-dominated regions.

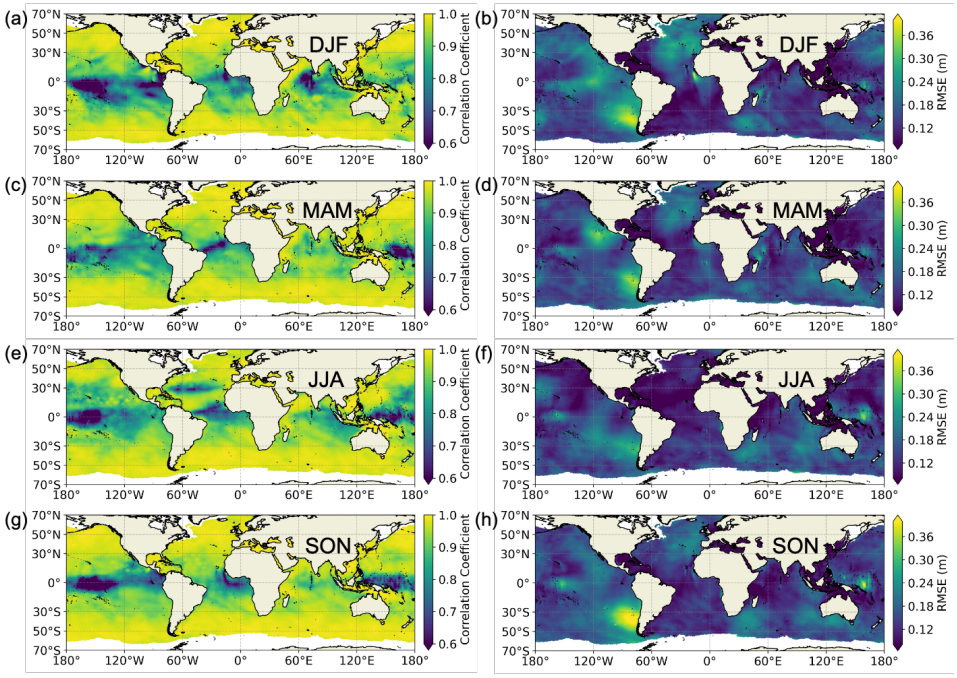

Figure R6. *Global distributions of (left) correlation coefficients and (right) RMSEs between AI model outputs and ERA5 for different seasons: (a,b) DJF, (c,d) MAM, (e,f) JJA, (g,h) SON.*

3. By focusing on analysing results after the errors stabilize, do the authors imply that this AI SWH model is more suitable for wave forecast beyond 10 days (240 hrs) without data assimilation and beyond 3-4 days with data assimilation?

No, we are not implying that the AI SWH model is more reliable when the errors become stable.

It is noted that while one application of our AI SWH model is SWH forecast, it is essential to emphasize that wave models—whether numerical or statistical—are not limited to forecasting wave conditions over a few days. Besides, the performance of wave forecasting, rely not only on the performance of wave model, but also on the accuracy of wind forecasting.

Here, the reason for focusing on the results after the errors stabilize is to demonstrate that the error of the AI rolling model does not accumulate indefinitely if the model is driven by high-quality forcing fields. More importantly, after reaching a stable state, the AI model achieves accuracy comparable to state-of-the-art numerical wave models, demonstrate its usability. With available observational data, data assimilation helps the model stabilize more quickly and achieve better results. This results indicate that such an AI wave model can be used for long-term hindcasts, projections, and rolling forecasts of SWH.

For real-world forecast/hindcast problem, it is noted there cannot be "perfect" initial field. The initial field for each forecast cycle derives from either the prior forecast field or analysis field. Consequently, during rolling forecasts, the model's initial conditions at every time step typically reach a stable error state. Given these initial fields, if the AI model driven by high-quality future wind fields (e.g., analyzed wind fields), the model would maintain stable error characteristics – this is the rationale for SWH hindcast, both for our AI model and NWMs. However, in operational wave forecasting, the driving wind fields themselves accumulate increasing errors with forecast lead time, inevitably leading to progressive degradation of wave forecast quality - an inherent limitation for all wave forecasting, also for both AI models or NWMs.

4. Although the authors acknowledged that this paper does not compare with in-situ observations, to showcase the effectiveness of this AI model, I think it can still be worthwhile to compare the AI SWH model, WW3-ST6 hindcast, and ERA5 reanalysis, against a few in-situ buoy observations in the manner of a short time series at some key locations (e.g., some key swell-dominated locations versus wind-sea dominated locations) or weather conditions (e.g., westerlies or more uniform wind conditions versus tropical or extra-tropical cyclones).

We sincerely appreciate the reviewer's valuable suggestion. In response, we have conducted a comparison between different models and NDBC buoy observations, following the same evaluation method used for CCI-sea state altimeter data. The results demonstrate consistency

with the comparisons using CCI-sea state altimeter data. For the AI rolling model without data assimilation, once the simulation stabilizes after approximately 240 hours of rolling inference, its performance is comparable to the state-of-the-art NWM, WW3-ST6. This further validates the effectiveness of our AI model. In contrast, for the models that used assimilation, there was a substantial improvement in all error metrics, demonstrating the effectiveness of assimilation in AI modeling. The following four figures have now been included in the revised manuscript and Supporting Information.

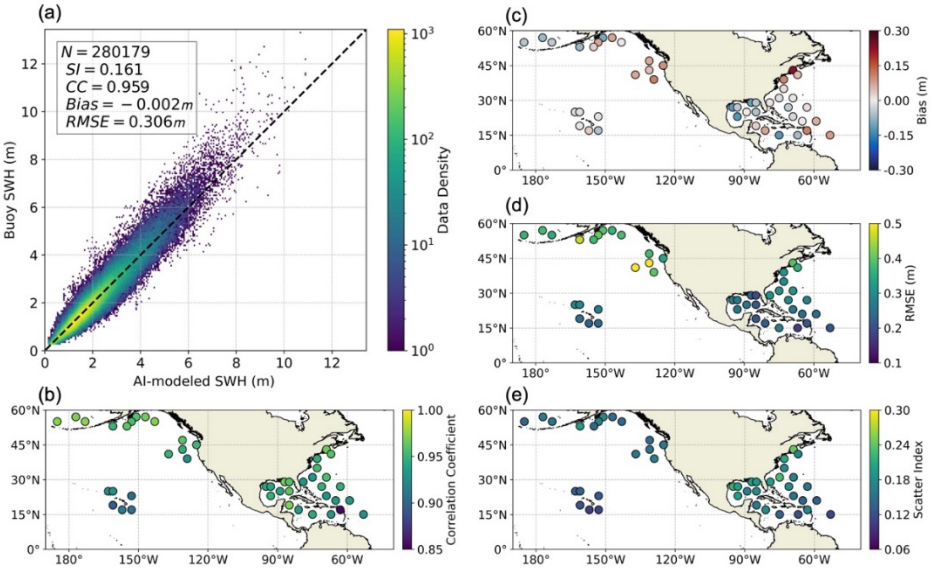

Figure R7. *The comparison between SWHs from the AI model and NDBC Buoy in 2020 for global ocean. (a) The scatter plot between the SWHs from the two datasets. (b-e) The spatial distributions of CC, bias, RMSE, and SI, respectively.*

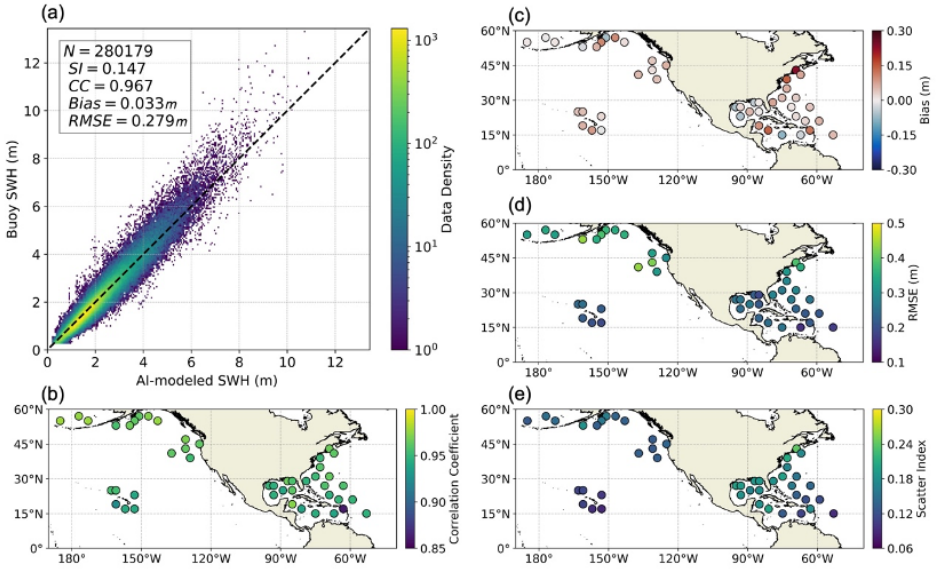

Figure R8. *The same as Figure R7, but the AI model has assimilated the data from CCI-Sea State every six hours.*

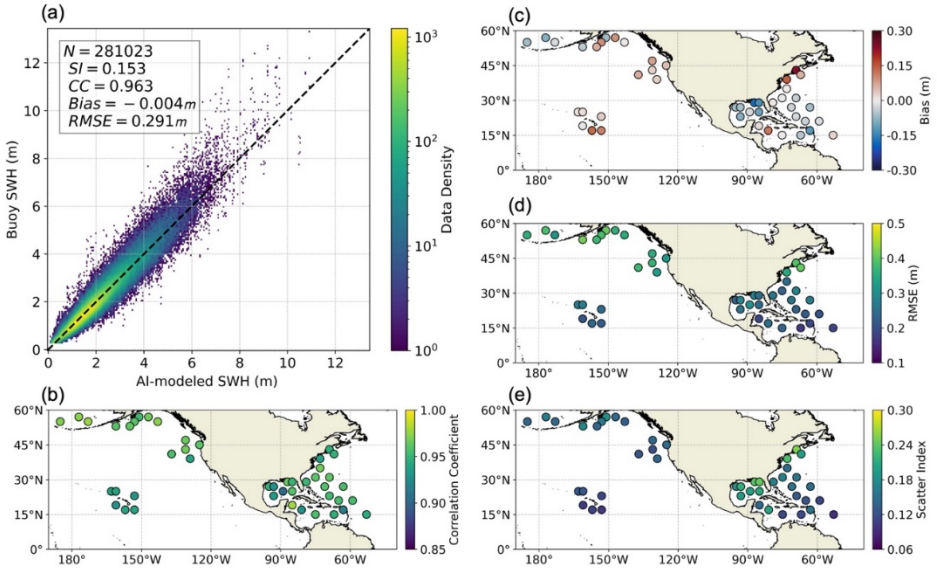

Figure R9. *The same as Figure R7, but the comparison is between the WW3-ST6 and NDBC Buoy.*

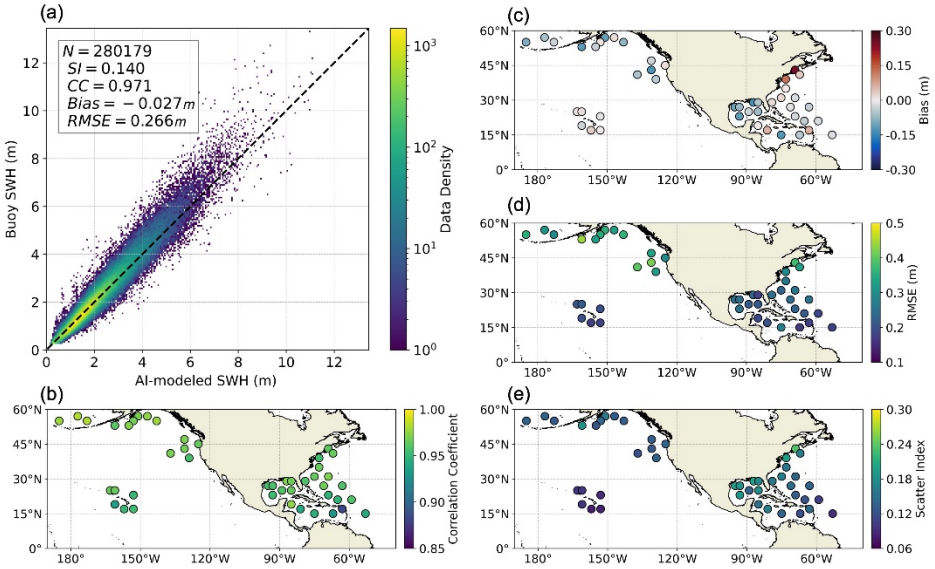

Figure R10. *The same as Figure R7, but the comparison is between the ERA5 and NDBC Buoy.*

Discussion:

It will be helpful if the authors can be more specific about the suitable applications with the AI SWH model. (e.g., time scales of the operational wave forecast, locations, seasons etc.)

We have put the discussions of the potential applications of the AI model in Section 5 (Concluding Remarks). Which reads: "An important advantage of the AI SWH model proposed here is its low computational cost compared to traditional NWMs. For example, on a personal laptop equipped with a single RTX 3060 GPU, the AI model can perform a 1-year global SWH rolling simulation at a resolution of 0.5° × 0.5° × 1h in just 10 minutes. In contrast, traditional NWMs, such as the WAVEWATCH III model, typically require several days to complete a simulation with the same output, even on supercomputing facilities. This makes the AI model particularly valuable in time-sensitive and resource-constrained scenarios, where it can be used as a surrogate for the NWMs. One potential application of this model is ensemble modeling,

both in operational wave forecasting and wave climate studies. In these applications, it is challenging to run NWMs multiple times using wind fields from different ensemble members of weather forecast models (for wave forecasting) or of various climate scenarios for long-term projection (for wave climate projection) due to the limitation of computational resources. In contrast, these tasks can be efficiently completed using the AI model, even on a standard laptop." These can be regarded as the potential applications of all AI wave models.

Besides, in the last paragraph, we wrote that: "We have demonstrated that the current SWH field and the wind field at the next time step are minimum requirements for the inputs of an AI SWH model. Such simplicity of model inputs and outputs makes this model a potential baseline for AI-based modeling of global SWH." This can be regarded as a special application for this specific model.

Regarding the time scales of the operational wave forecast, it is dependent on the performance of forecasting wind fields, while the ability of weather forecast is clearly beyond the scope of this study. Regarding the locations and seasons, as discussed in the manuscript, this AI model is more suited for wind-sea dominated conditions. Therefore, if a region has active wind fields (in some seasons), it will be suited for the application of the AI model (in the season that wind fields are active).