Response to Reviewer#2 Comments

The manuscript uses machine learning methods to establish relationships between tide gauge measurements and several atmospheric and oceanic variables, generating a global coastal storm surge dataset at 10 km spatial resolution. Overall, the generated dataset is of substantial application value, and the validation results show strong performance, particularly in the reconstruction of extreme values—a known challenge for AI models. The topic aligns well with the aims of ESSD. However, there are several key areas that require attention to ensure the manuscript is clear, methodologically sound, and accessible to readers.

Majors:

Point 1: The discussion of previous studies in the introduction lacks depth. The authors list previous studies without effectively explaining how the current work advances the field. To strengthen this section, the introduction should focus more on the existing gaps in storm surge modeling and how the proposed dataset addresses those shortcomings. The classification of storm surge research is overly simplified. The four categories mentioned in the second paragraph overlap and include one another. Moreover, the machine learning approach presented in this paper is described as separate from AI-based methods, though it clearly falls within that domain as a regression model. The difference between this approach and single-site models is primarily in the inputs used, such as geographic and temporal variables, but the fundamental methodology remains similar. A more refined categorization would provide better context for the reader.

Response:

Thanks for your constructive suggestions. In this version, we deleted the second paragraph and rewrote the introduction in Lines 31-84 to make it focus on the existing gaps in storm surge datasets and how our method can fix the gaps. The logic is from tide gauge observations to numerical models, then to data-driven models:

High-frequency (at least hourly), sufficient spatial coverage, and long-term records are the basis for SS analysis. To date, tide gauges (TGs) are the most reliable source of coastal sea-level observations (Marcos et al., 2019). However, their distribution is sparse and uneven. For example, as the most complete high-frequency TG collection currently, though the Global Extreme Sea Level Analysis version 3 (GESLA-3) dataset included 5,119 stations around the world, most of them were distributed in North America, Europe, Japan, and Australia (Haigh et al., 2023).

Interpolating TG observations among different stations cannot accurately capture the variabilities of SSs (Muis et al., 2016) since they are affected by many factors, such as storminess, coastline shape, and bathymetry (Resio and Westerink, 2008). This always limits in-depth analysis of the spatial characteristics of SSs from TG records directly, especially on a global or quasi-global scale. In addition, though some of the oldest TG stations can date back to the eighteenth century, only ~10% (554 stations) of TG records in the GESLA-3 dataset were longer than 50 years, which makes it difficult to obtain more detailed long-term variations in SSs.

Numerical models can provide simulated data with better spatial coverage by resolving coastal physical processes inducing SSs (Muis et al., 2016, 2023; Lockwood et al., 2024). A common limitation of numerical models is that they require accurate and high-resolution bathymetric data for sufficiently precise SS estimations since SSs are significantly affected by water depth in shallow water (Resio and Westerink, 2008). However, such bathymetric data is often unavailable in nearshore areas (Cid et al., 2018). In addition, in global or quasi-global SS simulations, the coastal grid resolution of numerical models is usually set to several kilometers to balance the computational complexity (Muis et al., 2020; Mentaschi et al., 2023), which means that nearshore physical features with a spatial scale smaller than this resolution cannot be sufficiently simulated (Parker et al., 2023), and hence affecting the SS precision. Meanwhile, the computational efficiency of global numerical models tends to affect the length of simulated SSs (Muis et al., 2019). For instance, the state-of-the-art Global Tide and Surge Model (GTSM), though its outputs have been widely used in relevant studies (Kirezci et al., 2020; Dullaart et al., 2021; Fang et al., 2021; Yang et al., 2024b), its simulations spaned only the most recent decades from 1979 to 2018 (Muis et al., 2020). This imposed limitations on studies requiring long-term SS records.

Unlike numerical models, data-driven models do not need to resolve coastal physical processes. They obtain the statistical relationship between SSs (predictand) and relevant atmospheric factors (predictor) through multiple linear regression (Cid et al., 2018) or artificial intelligence (Bruneau et al., 2020). Therefore, the precision of data-driven models is unaffected by bathymetric data and grid resolution. In addition, long-term SSs can be reconstructed efficiently after the statistical relationship is established (Tadesse et al., 2020). However, the commonly used single-site modeling framework for data-driven models heavily relies on TGs; it must establish independent relationships for every TG site by site (Cid et al., 2017; Bruneau et al., 2020; Tiggeloven et al., 2021) and cannot provide any SS information at ungauged coastal locations. For

example, the Global Storm Surge Reconstruction (GSSR) database, the only publicly released global SS dataset from the data-driven model, provided SS reconstructions at 882 points globally going as far back as 1836, which benefited the research on long-term trend analysis of SSs (Tadesse and Wahl, 2021). However, it cannot address issues caused by the sparseness and uneven distribution of TG stations. Some studies replaced TG observations with numerical SS simulations to train the data-driven model (so-called 'surrogate model') (Lee et al., 2021; Ayyad et al., 2022; Lockwood et al., 2022). This combination improved the spatial resolution, but numerical models' precision limitations were also transferred to the surrogate model. Moreover, in theory, surrogate models cannot be better than numerical models compared to TG observations. Yang et al. (2023) proposed a novel all-site modeling (ASM) framework, which allowed the data-driven model to reconstruct high spatial-coverage SSs in research areas by learning from TG observations (without SS simulations from numerical models). Although single-site modeling and ASM belong to the data-driven model, their basic ideas differ. The former presumes SS observations at different TGs are independent. Therefore, the relationship between predictors and SSs needs to be learned for every TG site by site; this relationship is unsuitable for other locations. In contrast, the latter believes there is a universal connection between SSs at different TGs, so all available TGs within the research area can be pooled into one model to learn the only regional relationship between predictors and SSs. This essential difference enables the ASM framework to reconstruct SSs at any coastal point in the research area. In addition, the study has shown that ASM's precision is better than single-site modeling's (Yang et al., 2023).

High spatiotemporal resolution and sufficiently long SS dataset is the basis for analyzing this disaster. However, the existing SS datasets, whether from TG observations, numerical model simulations, or data-driven reconstructions, cannot fulfill all three demands simultaneously on a global or quasi-global scale. The ASM provides an opportunity to fix this gap. This research used it to establish a SS data-driven model in coastal areas within ~45°S to ~45°N, which are severely affected by SSs since most destructive tropical and extratropical cyclones occur here (Knapp et al., 2010). After precision assessment by comparing it with TG observations and the numerical model GTSM, we released, for the first time, a long-term (> 80 years from 1940 to 2020) quasi-global hourly SS dataset reconstructed from the data-driven model with high spatial resolution (10 km along the coastline). We hope this dataset, the ASM-SS (all-site modeling storm surge), can provide possible alternative support for coastal communities to deepen our understanding of SSs and ESLs.

Point 2: The description of the model's methodology lacks sufficient detail on its innovations. For instance, the choice of specific atmospheric and oceanic variables from ERA5 should be justified, and the process of integrating geographical and temporal variables requires further explanation. How were these inputs pre-processed to allow for prediction across any coastal location or time? This is a key aspect of the model and should be clarified. Although more detailed explanations may have been presented in the authors' previous publications, it is still important to concisely convey these methodological details in this data-focused paper to ensure readers can fully understand the process without referring to other sources.

Response:

We are sorry for any difficulties in understanding. We added more details of our modeling framework in this version (Lines 135-150), hoping it can be clearer and more understandable to readers:

Full details of the ASM can be found at Yang et al.(2023). Here, a conceptual description of it is provided. As mentioned in the introduction, though single-site modeling and ASM belong to the data-driven model, their basic ideas are different. For example, assuming there are n available TGs within 45°S to 45°N. The single-site modeling presumes SS observations at n TGs are independent; each site needs to build a separate data-driven model to learn the relationship between predictors and SSs at that station. In this case, n single-site modeling data-driven models are established, and they cannot reconstruct SSs for locations other than TG stations. Unlike single-site modeling, the ASM believes a general connection exists between SSs at n TGs within the research area. Namely, there is a unique regional relationship between predictors and SSs, and all TGs follow this relationship. Therefore, predictors and SSs at n available TGs can be pooled into one ASM data-driven model. After learning the only regional relationship through adequate training, this ASM model can be used to reconstruct SSs at any gauged or ungauged coastal point within the research area by inputting relevant predictors. The following are the modeling processes:

(1) Obtaining predictors. Four atmospheric data (mslp, u10, v10, and t2m) for each TG station are extracted from the ERA5 dataset through linear interpolation. Changes in sea level pressure and wind are the main factors in generating SSs (Woodworth et al., 2019); adding temperature variations considers the effects of thermal expansion and contraction. Meanwhile, following Yang et al.(2023) and Yang et al. (2024a), another three variables (longitude, latitude, and timestamp) are considered since geographical locations and record lengths of TGs are different. Hence, the predictor matrix for each

TG consists of 7 columns: mslp, u10, v10, t2m, longitude, latitude, and time;

Point 3: One of the key strengths of the model is its superior performance in predicting extreme storm surge events compared to numerical models. However, the reasons behind this superior performance are not fully explored. A deeper analysis of why the machine learning model performs better than numerical models in extreme cases, particularly considering that AI models often struggle with extremes, would add significant value.

Response:

Thanks for your suggestion. In this version, we discussed the possible reason in Lines 228-234 after the comparison between our data-driven mode and numerical model in section 3.2:

The reason why ASM outperforms GTSM can be attributed to two main aspects. For the global numerical model GTSM, as mentioned in the introduction, the accuracy and spatial resolution of bathymetric data in the nearshore area limits the precision of SSs. Meanwhile, the grid with a resolution of several kilometers affects the effective simulation of small-scale physical factors. For the ASM data-driven model, the training process is based on TG observations. TGs are the most accurate source for sea level monitoring, and their records can be considered to include effects from all spatial-scale physical processes. In addition, the machine learning method XGBoost is a residual model that pays more attention to where residual errors are significant, which also benefits the estimation of extreme SSs.

Point 4: While the manuscript provides a thorough discussion of the spatial performance of the dataset, it lacks an analysis of the model's temporal performance. How does the model perform over the 1940–2020 period? Are there periods when the model is more or less accurate? Providing this temporal analysis would add an important dimension to the validation results.

Response:

Thanks for your professional suggestion. In this version, we added the temporal variation analysis of our model's precision from 1940 to 2020 every 10 years in Lines 190-206:

It is necessary to evaluate temporal variations in reconstructed SSs further since their length is over 80 years, during which the number of TG stations and the quality of atmospheric data have changed. As shown in Fig. 4, the precision of ASM model at

TGs in each sub-region was calculated every 10 years (excluding TGs with less than one year of data in a given decade). Results indicate that the overall precision (i.e., for ALL TGs) of entire surges and 95th extremes gradually increased from 1940 to 2020. Possible reasons are, on the one hand, the increase of TGs in recent decades provided more predictand features; on the other hand, the optimization of ERA5 atmospheric data (predictor) contained more detailed tropical and extratropical cyclone information. At the regional scale, for entire surges, Fig. 4(a) indicates that except for SWA (CORR decreases) and WAS (CORR remains unchanged), CORRs of other sub-regions present an upward trend; Fig. 4(b) shows the RMSE in SES increases, while RMSEs in other regions decrease; Fig. 4(c) gives that MBs of sub-regions have been gradually optimized (excluding WAS). For 95th extremes, in terms of CORR (Fig. 4(d)), WEU, NAF, WNA, ENA, EAS, NOC, and SOC show an upward trend, whereas there is no obvious pattern in other regions; for RMSE (Fig. 4(e)), ER, SEA, and SES present an increasing trend, other regions decrease; for MB (Fig. 4(f)), the underestimation of SSs in ER and SAS rises, and there is no noticeable change in WNA and SES. MBs in WEU, NAF, ENA, WAS, EAS, NOC, and SOC are optimized, while there is no clear pattern in SWA, SEA, CA, and SWS.

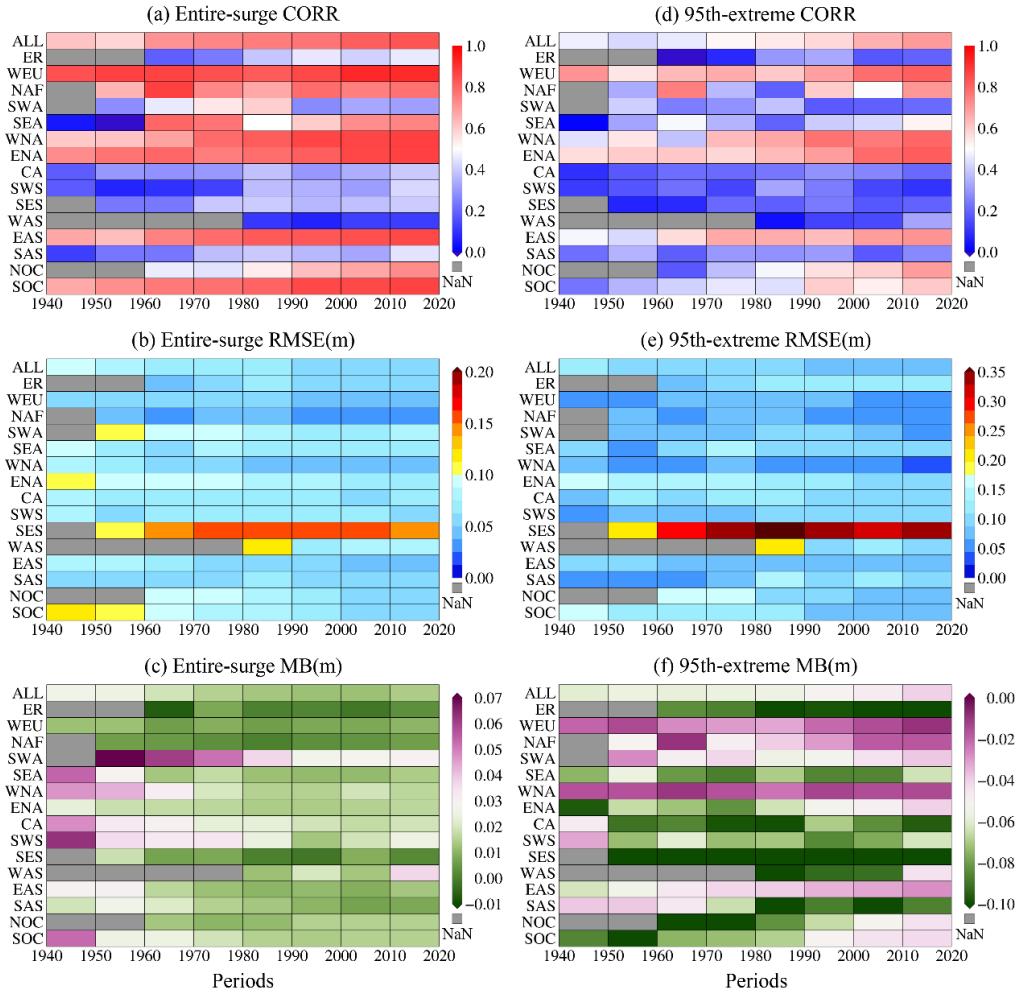

Figure 4: Temporal variations of the ASM model's precision at tide gauges from 1940 to 2020. (a-c) Entire surge evaluation statistics for different regions every 10 years; (d-f) 95th extreme evaluation statistics for different regions every 10 years

Point 5: Figure 1 shows several tide gauge stations in South America and West Africa with long records, yet these regions are not featured in the validation results. The authors should explain why results from these areas were excluded from the analysis.

Response:

Apologies for this confusion. Our initial logic was to analyze the areas mainly affected by tropical or extratropical cyclones in the main text (since there are almost no tropical or extratropical cyclones in the equatorial region, the South Atlantic and the southeastern Pacific), and put the evaluation results of the entire area in the appendix. However, this did not seem to be a suitable way. Therefore, in this version, we deleted the appendix and moved the assessment of the entire domain into the main text for discussion. All relevant figures were updated:

As shown in Fig. 3, we divided the research area into fifteen sub-regions (ER: the

equatorial region, WEU: Western Europe, NAF: Northern Africa, SWA: Southwestern Africa, SEA: Southeastern Africa, WNA: Western North America, ENA: Eastern North America, CA: Central America, SWS: Southwestern South America, SES: Southeastern South America, WAS: Western Asia, EAS: Eastern Asia, SAS: Southern Asia, NOC: Northern Oceania, and SOC: Southern Oceania) for more detailed assessment information. Note that the equatorial region (~6°S to ~6°N) was separated as an independent area since it has almost no tropical or extratropical cyclones.

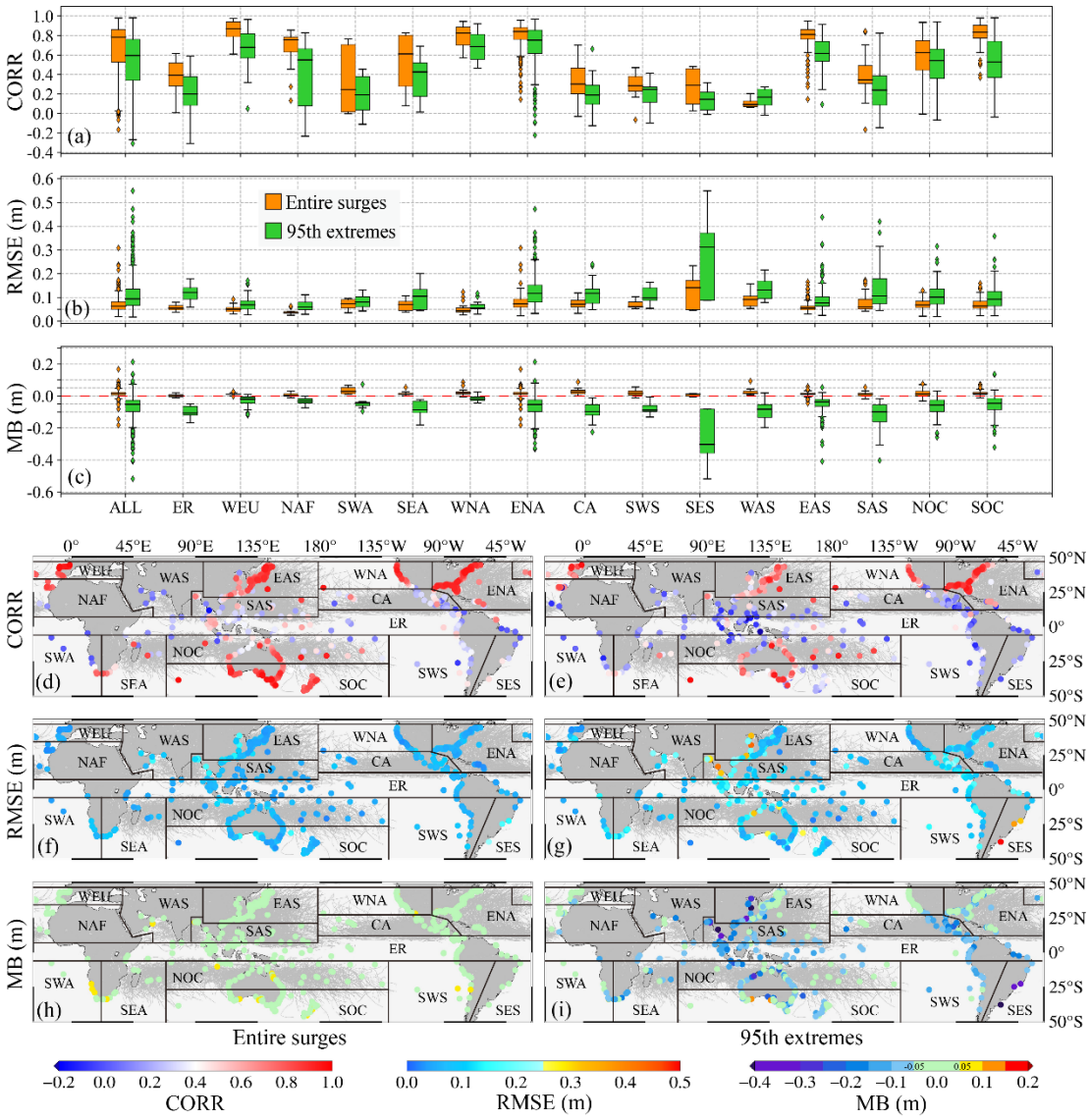

Figure 3: ASM model evaluation at tide gauges from 1940 to 2020. (a-c) Entire surge and 95th extreme evaluation statistics for different regions; (d-i) Distributions of evaluation metrics. Gray lines are tropical cyclone paths.

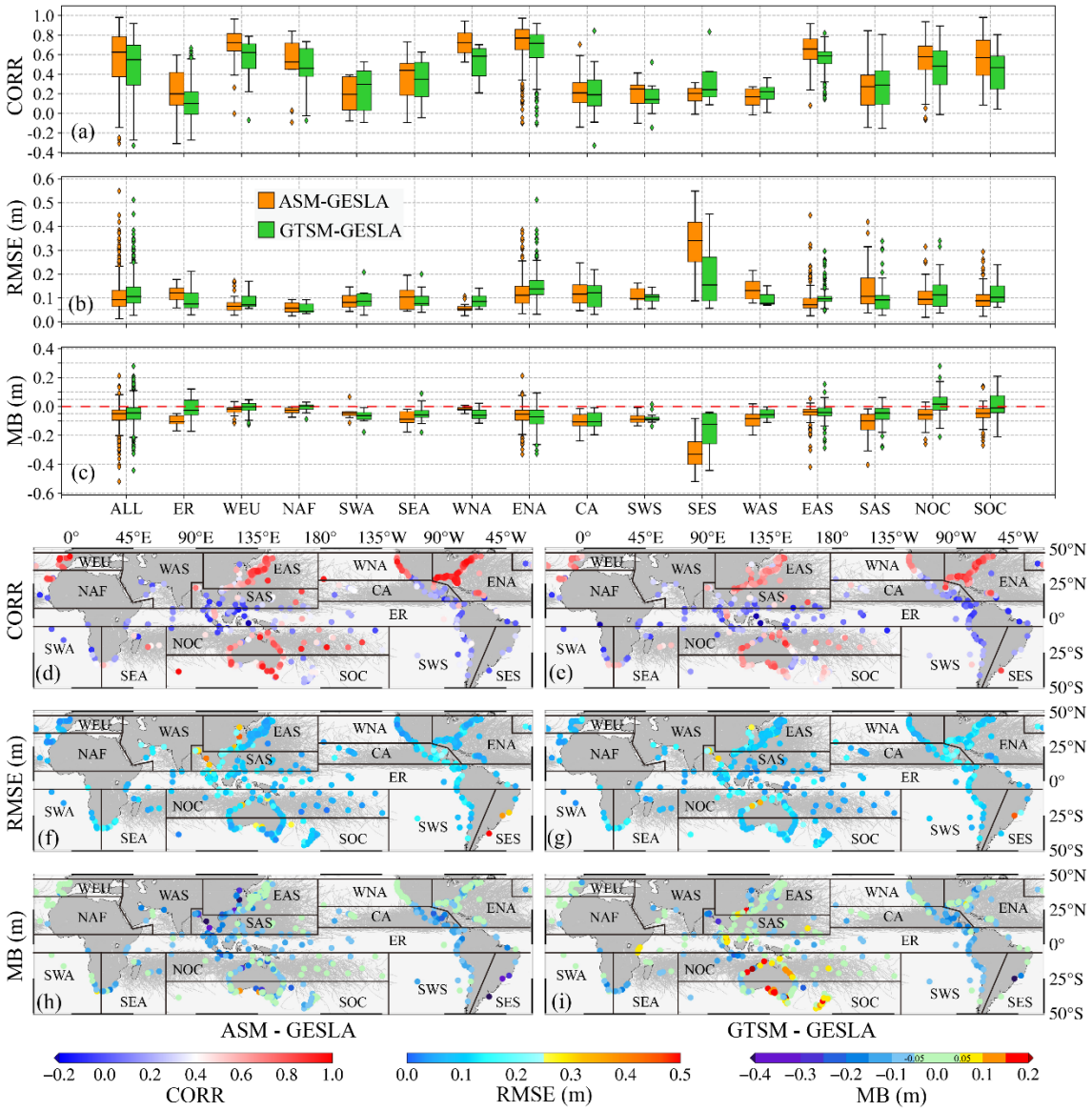

Figure 5: ASM model comparison with the numerical model at tide gauges from 1979 to 2018. (a-c) ASM and GTSM 95th extreme evaluation statistics for different regions; (d-i) Distributions of evaluation metrics. Gray lines are tropical cyclone paths.

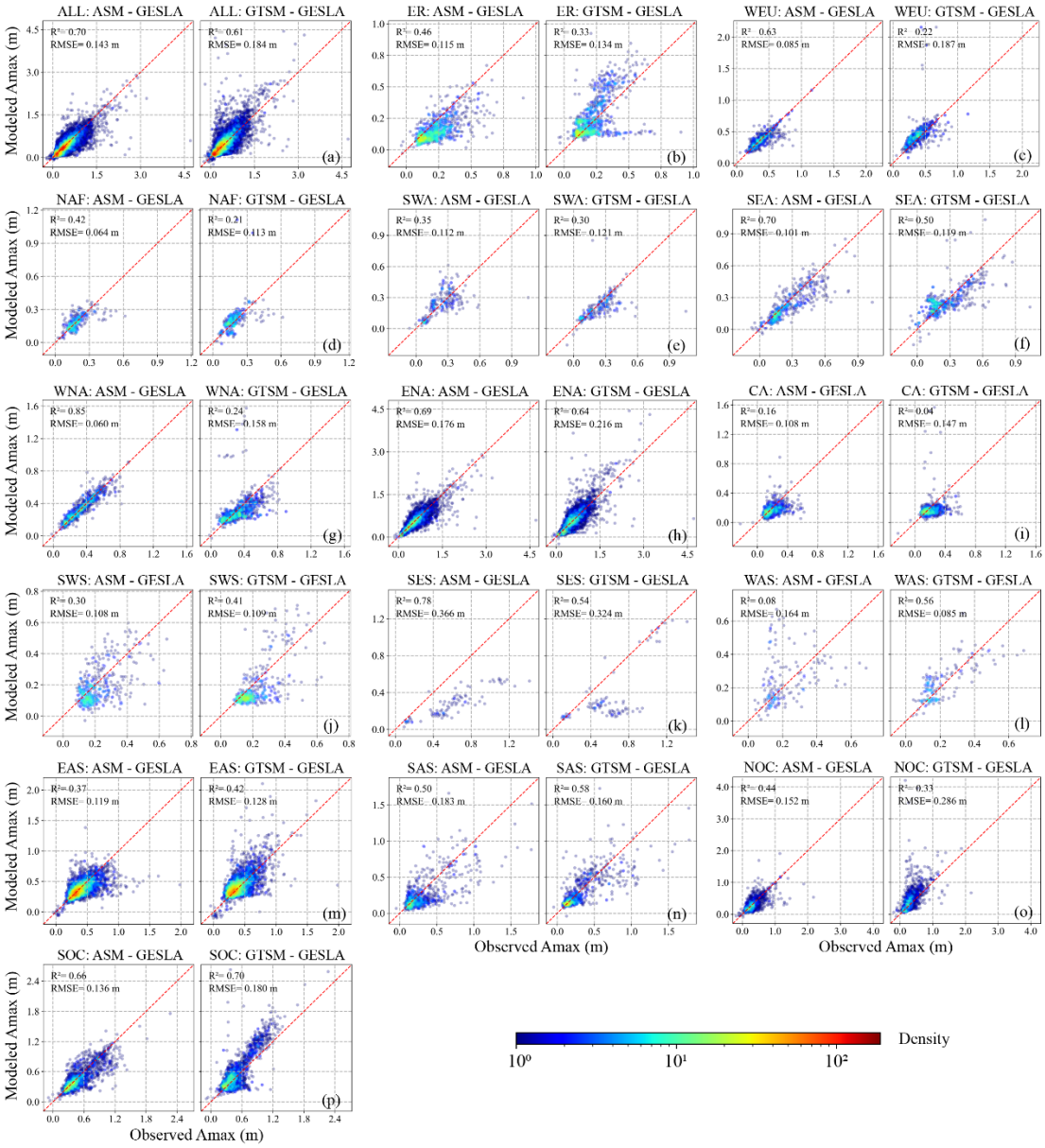

Figure 6: Scatter density plots of ASM and GTSM annual maxima (Amax) compared with tide gauge observations in different regions. The data for tide gauges were combined. The red dotted line indicates the perfect fit line.

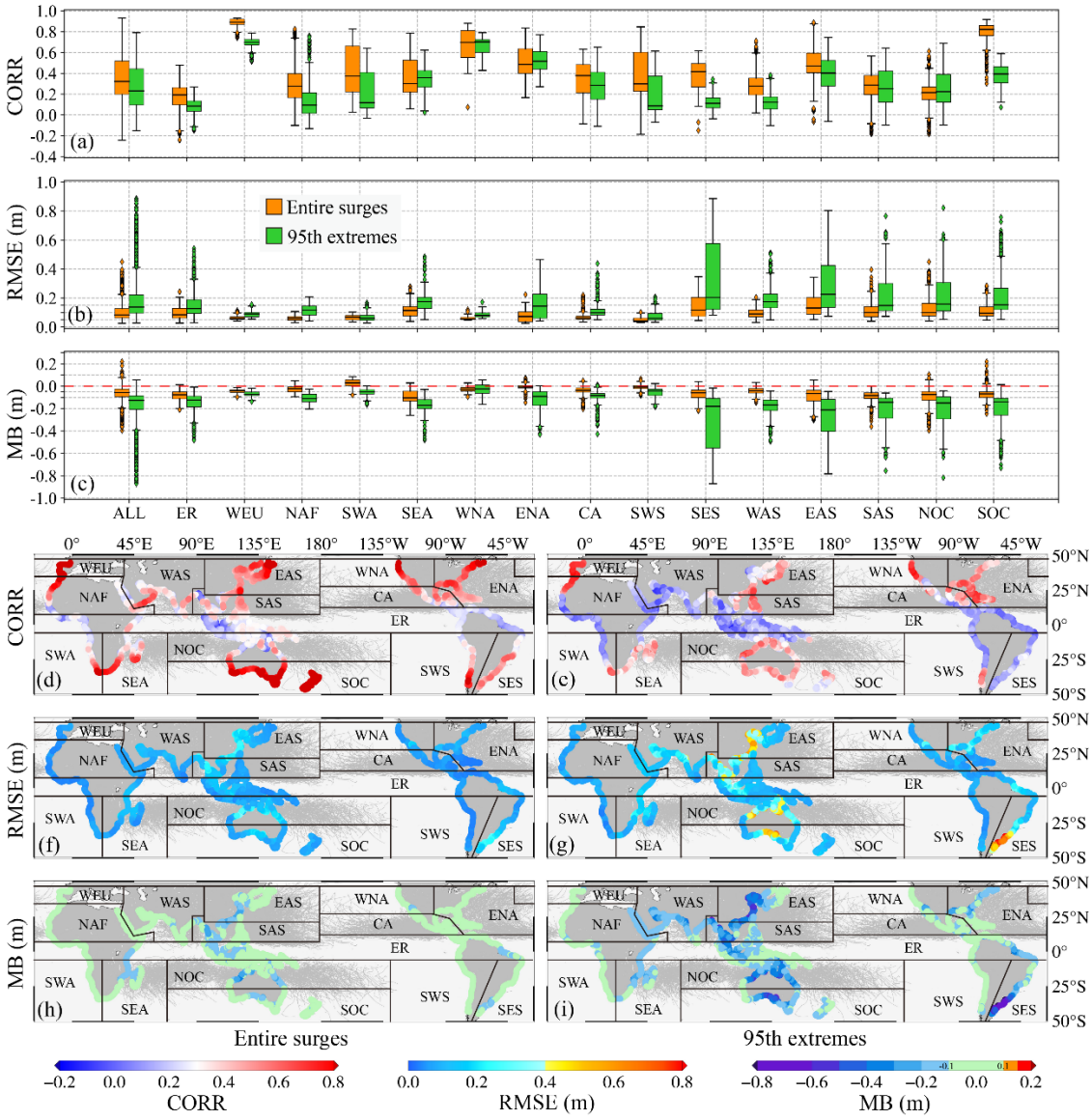

Figure 7: Differences between ASM and GTSM at the coastal scale from 1979 to 2018. (a-c) Comparison statistics between ASM and GTSM modeled entire surges and 95th extremes for different regions; (d-i) Distributions of comparison metrics. Gray lines are tropical cyclone paths.

Point 6: The manuscript suffers from imprecise language and grammatical errors. Phrases like "coastline having complicated shapes" (line 41) and "internal climate variability" (line 49) are vague and not commonly used in geoscience literature. Additionally, phrases such as "numerical models are based on shallow water equations" (line 65) overly simplify the complexity of these models. Grammatical issues such as "until now" (line 9) and "will" (line 89) create ambiguity and should be corrected for clarity. Moreover, the manuscript contains an excessive number of speculative terms such as "some," "might," "may," and "slightly better." Scientific writing should avoid

this level of uncertainty when possible, and more precise language should be used. Where quantifiable data are available, the authors should provide specific numbers to reduce ambiguity.

Response:

We are sorry for our imprecise language and grammatical errors. Phrases "coastline having complicated shapes" and "internal climate variability" were deleted since we rewrote the introduction; "numerical models are based on shallow water equations" was replaced with "…by resolving coastal physical processes inducing SSs" in Line 41. In addition, we carefully revised the tense issues and reduced the use of speculative terms in this version, hoping it can be more readable and precise.

Point 7: The authors should ensure that the data description fully complies with the journal's requirements. Additional details about the structure and usage of the dataset may be necessary for ESSD's standards.

Response:

Thanks for reminding. We added more details about the dataset in Lines 262-271:

The ASM-SS quasi-global storm surge dataset was generated from the ASM data-driven model established in section 3.3. The dataset is available at https://doi.org/10.5281/zenodo.13293595 (Yang et al., 2024a) as NetCDF files month by month from 1940 to 2020. Each file includes five parameters: longitude, latitude, nodes, time, and surge level. Longitude and latitude are the location information of nodes in degree; the unit of time is accumulated hours since 1900-01-01 00:00:00; surge levels are given in meters. Users can use longitude, latitude, and time as keywords to select surge levels at nodes of interest within a target period. In addition, the spatial resolution of nodes is 10 km along the coastline (as shown in Fig. 2). Since the sea surface varies rapidly during tropical cyclones, the temporal resolution of surge levels is set to hourly. Though this temporal resolution increases the data volume, it can provide sufficient information for users who want to analyze high-frequency variations of storm surges during extreme events.

Minors:

Point 8: The choice of an hourly temporal resolution for the dataset is not fully explained. The authors should provide a rationale for this decision, especially considering the implications for data volume and usability.

Response:

Thanks for your recommendation. The reason was added in Lines 268-271:

Since the sea surface varies rapidly during tropical cyclones, the temporal resolution of surge levels is set to hourly. Though this temporal resolution increases the data volume, it can provide sufficient information for users who want to analyze high-frequency variations of storm surges during extreme events.

Point 9: The mention of "small phase shifts" (line 120) lacks context. The origin of these phase shifts and their impact on the results should be discussed in detail.

Response:

Thanks for your suggestion. Previous research has discussed this issue. For example, Horsburgh and Wilson (2007) gave the following figure:

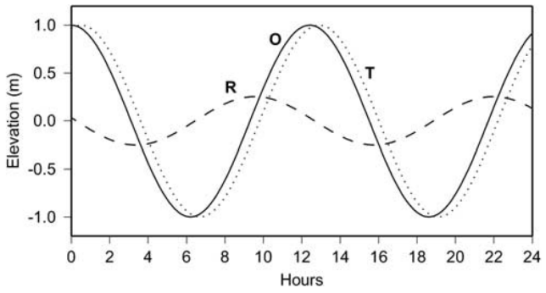

Figure 4. Schematic diagram of a sinusoid whose phase is altered but whose frequency and amplitude remain unaltered. The solid line (O) represents observations, the dotted line represents tidal predictions (T) and the dashed line represents the residual obtained via subtraction (R).

(Horsburgh, K. J. and Wilson, C.: Tide-surge interaction and its role in the distribution of surge residuals in the North Sea, J. Geophys. Res., 112, 2006JC004033, https://doi.org/10.1029/2006JC004033, 2007.)

In this version, we presented the reason in Lines 110-113:

(5) Finally, a 12-hour moving average was applied to SS data to limit possible remaining tidal signals (Tiggeloven et al., 2021; Yang et al., 2023), which are generally generated by small phase shifts in predicted tides due to the difficulty of obtaining perfect and completely accurate estimates through harmonic analysis (Horsburgh and Wilson, 2007).

Point 10: Units such as cm/m should be standardized across the manuscript. Similarly, decimal precision should be consistent for a more professional and coherent presentation of the data.

Response:

Thanks for your professional suggestion. We standardized the units of RMSE and MB to meters with three decimal places. The precision of CORR was set to two decimal places:

Lines 14-16: the precision of this model (medians of correlation coefficients, root mean square errors, and mean biases are 0.63, 0.093 m, and -0.049 m, respectively) is better than that of the state-of-the-art global hydrodynamic model (medians are 0.55, 0.106 m, and -0.044 m);

Lines 182-184: Figure 3(a-c) show that on a quasi-global scale (i.e., for ALL TGs), the median CORR of the entire time series of surges is 0.78, RMSE is 0.062m, and MB is 0.014m. In comparison, the reconstruction precision for extreme events (>95th percentile) is lower: CORR is 0.59, RMSE is 0.094m, and MB is -0.052m.

Lines 215-217: ASM (medians of CORRs, RMSEs, and MBs for 95th extremes are 0.63, 0.093 m, and -0.049 m, respectively) outperforms the numerical model GTSM (medians are 0.55, 0.106 m, and -0.044 m).

Lines 249-251: there are noticeable differences between ASM and GTSM. On the quasi-global scale, medians of CORRs, RMSEs, and MBs of the entire surges (95th extremes) between them are 0.32 (0.23), 0.084 m (0.138 m), and -0.056 m (-0.126 m),

Point 11: The gray lines in the figures (presumed to be tropical cyclone paths) should be explicitly described, and their inclusion justified. What purpose do these lines serve, and how do they enhance the understanding of the storm surge dataset?

Response:

Thanks for reminding. Since not all areas between 45°S to 45°N are affected by tropical cyclones (for example, the equatorial region, the South Atlantic and the southeastern Pacific), mapping the tropical cyclone paths can facilitate the division of sub-regions and highlight their differences. We added relevant descriptions in this version:

Lines 177-178: …Note that the equatorial region (~6°S to ~6°N) was separated as an independent area since it has almost no tropical or extratropical cyclones.

Lines 180-181: Figure 3: ASM model evaluation at tide gauges from 1940 to 2020. (a-c) Entire surge and 95th extreme evaluation statistics for different regions; (d-i) Distributions of evaluation metrics. Gray lines are tropical cyclone paths.

Lines 213-214: Figure 5: ASM model comparison with the numerical model at tide gauges from 1979 to 2018. (a-c) ASM and GTSM 95th extreme evaluation statistics for different regions; (d-i) Distributions of evaluation metrics. Gray lines are tropical cyclone paths.

Lines 245-247: Figure 7: Differences between ASM and GTSM at the coastal scale from 1979 to 2018. (a-c) Comparison statistics between ASM and GTSM modeled entire surges and 95th

extremes for different regions; (d-i) Distributions of comparison metrics. Gray lines are tropical cyclone paths.

Point 12: The same color bar is used for multiple metrics, which can create confusion. I recommend using separate color bars for each metric to avoid misinterpretation.

Response:

Thanks for your recommendation. We used three color bars to show different metrics in this version, hoping it can be clearer. Taking Figure 3 as an example:

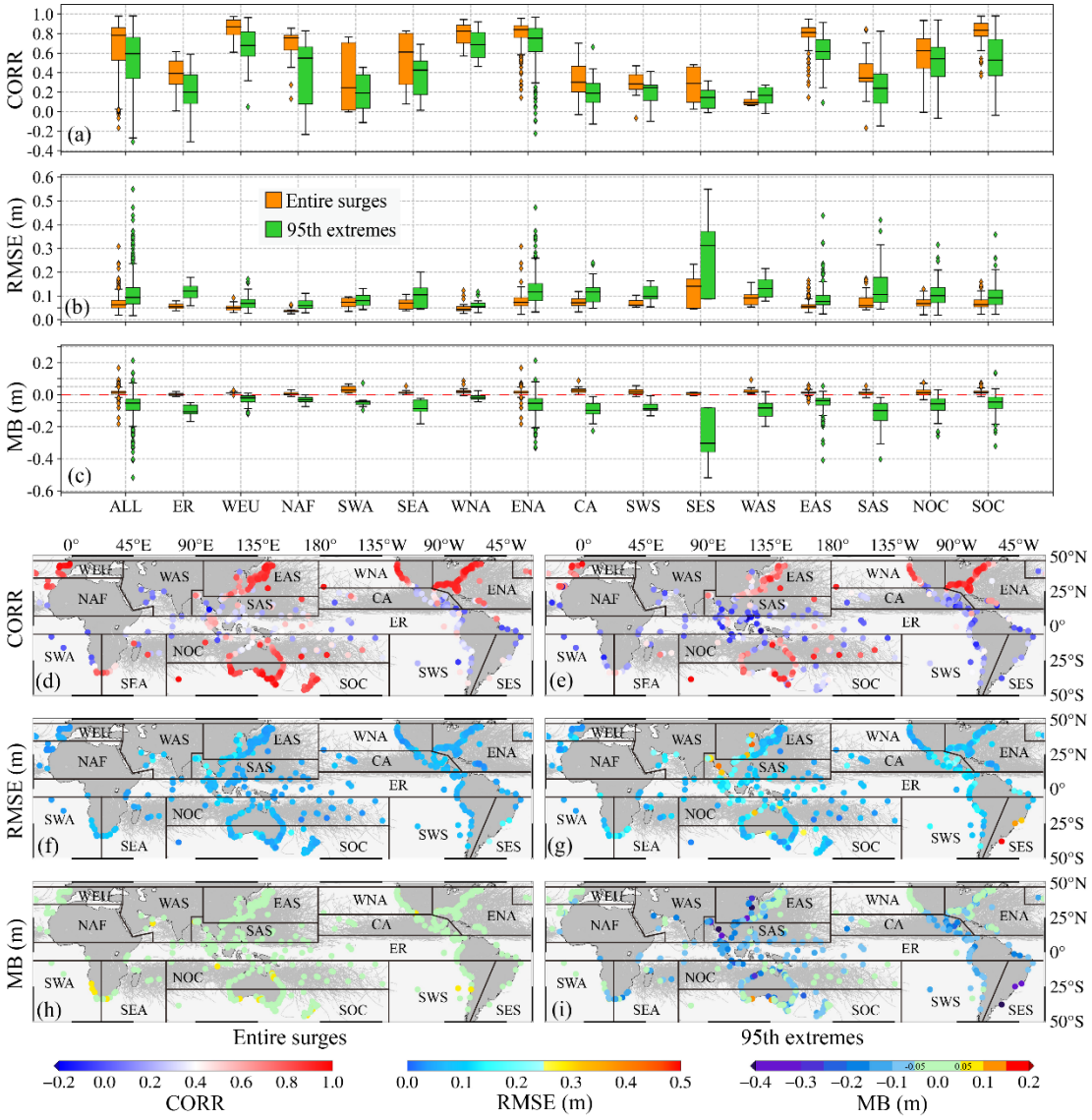

Figure 3: ASM model evaluation at tide gauges from 1940 to 2020. (a-c) Entire surge and 95th extreme evaluation statistics for different regions; (d-i) Distributions of evaluation metrics. Gray lines are tropical cyclone paths.

Point 13: The use of "surge" as a variable name in the NetCDF files is problematic, as

it refers to a physical phenomenon rather than a dataset variable. I recommend choosing a more precise name that clearly describes the data field.

Response:

Thanks for your recommendation. We replaced it with 'surge level'. The new NetCDF files were updated in the repository.

Point 14: Line 62, a space between "abovementioned".

Response:

Thanks for reminding. This word was not used in this version since we rewrote the introduction.

Another comment:

Point 15: The manuscript emphasizes the computational inefficiency of numerical models, but fails to acknowledge that AI models, particularly those involving extensive preprocessing, ground truth acquisition, and training, can also be computationally expensive. Large/big AI models often require substantial computing power. A more balanced comparison of the computational demands of AI models versus numerical models would provide a fairer perspective on the advantages and limitations of each approach.

Response:

Thanks for your constructive suggestion. As you mentioned in **Point 1**, this paper should focus on the gaps between existing storm surge models or datasets to highlight the advantages of our dataset; another reviewer holds the same opinion. Computational efficiency is not the most concerning matter for dataset users. They care more about what the new dataset can provide to their research. Therefore, we rewrote the introduction and deleted relevant descriptions in other sections.

Nevertheless, as you point out, with the expanding application scenarios of AI, finding a more objective way to evaluate its computational efficiency and resource cost is worth attention.

Summary:

Overall, this manuscript presents a highly valuable and timely contribution to the field of storm surge modeling. The application of machine learning to generate a global, high-resolution dataset fills an important gap in coastal hazard prediction, especially for regions lacking sufficient observational data. The dataset's strong performance in

reconstructing extreme values, combined with its spatial resolution, demonstrates its potential for numerous applications in coastal risk management and scientific research. While there are areas that could benefit from further clarification and refinement, particularly in terms of methodological transparency and computational comparisons, the work is commendable. It reflects a significant step forward in leveraging AI for oceanographic data analysis, and with some improvements, it will undoubtedly become a highly valuable resource for the community.