Assessing the storm surge model performance: What error indicators can measure the skill?

Rodrigo Campos-Caba[1], Lorenzo Mentaschi[1], Jacopo Alessandri[1], Paula Camus[2], Andrea Mazzino[3], Francesco Ferrari[3], Ivan Federico[4], Michalis Vousdoukas[5], and Massimo Tondello[6]

[1]Department of Physics and Astronomy, University of Bologna, Bologna, Italy
[2]Departamento de Ciencias y Técnicas del Agua y del Medio Ambiente, University of Cantabria, Santander, Spain
[3]Department of Civil, Chemical and Environmental Engineering. University of Genoa, Genoa, Italy
[4]Centro Euro-Mediterraneo sui Cambiamenti Climatici – Ocean Predictions and Applications, Lecce, Italy
[5]Joint Research Centre, European Commission, Seville, Spain
[6]HS Marine SrL, Noventa Padovana, Italy

Correspondence to: Rodrigo Campos-Caba (rodrigo.camposcaba@unibo.it)

1 Referee #1

General reply: Thank you for your thoughtful and constructive comments that helped us improve our manuscript. Below, we provide detailed responses to your specific comments and outline the corresponding revisions made to the manuscript.

<u>Specific comments</u>

1. LL59-66: "*Over recent years, unstructured grid models have increasingly emerged as alternatives to regular grids for large-scale simulations (e.g. Mentaschi et al., 2020; Muis et al., 2016; Vousdoukas et al., 2018; Fernández-Montblanc et al., 2020; Saillour et al., 2021; Wang et al., 2022; Zhang et al., 2023; Mentaschi et al., 2023), with established circulation unstructured models like […] Delft3D (Deltares: Delft, 2024), among others*."

The standard Delft3D model is actually based on structured grids. In this context, you should specifically refer to Delft3D-FM, which uses unstructured grids.

R: We thank the reviewer for pointing this out. We have corrected this in the revised manuscript, now specifically referring to Delft3D-FM at line 66.

2. LL142-144: "*Tides with hourly resolution from the Finite Element Solution (FES) 2014 (Lyard et al., 2021) were also included to account for the total sea level in the simulations*."

Were all tidal constituents from FES2014 included in the boundary forcing, or was just a selection of constituents used?

R: Thank you for your question. We used a selection of tidal constituents based on harmonic analysis at the relevant locations. To provide more detail, we have now specified the selected constituents on the revised manuscript, adding the following (lines 146-148): '*Specifically, the constituents included for the tide reconstruction are SA, SSA, O1, P1, S1, K1, N2, M2, MKS2, S2, R2, K2, M3, M4, and MS4, which were selected based on preliminary harmonic analysis applied to sea level observation data in the locations specified in Section 2.2.*'

3. LL206-207: "*Additionally, with the aim of considering the representation of extremes by the simulations, we introduce two new metrics based on customized versions of the Mean Absolute Deviation (MAD)*:"

a) Why don't you use the traditional MAD as a metric to assess your models' qualities, given that your new metrics are based on it? Does the traditional MAD already exhibit similar behaviour to MADp and MADc in identifying BC-UniGe as the best model? If it does, this would raise the question of why the new metrics are necessary.

R.a: Thank you for this insightful comment. We agree that including the traditional MAD metric is important for a comprehensive evaluation. In response, we have now included the traditional MAD in the revised manuscript as an evaluation metric. The mathematical expression is presented in Equation 7, and results are incorporated into the scatter plots (Figures 4 and 5) and radar charts (Figures 3 and 6). We have also added corresponding discussions on the results obtained with this metric. The added and/or modified lines in the revised version of the manuscript are the following:

Lines 230-234: '*Mean Absolute Deviation (MAD):*

$$\boldsymbol{MAD = \overline{|S - O|}} \qquad (7)$$

***A value closer to one indicates a better performance.*'**

Lines 283-284: '*The performance evaluation shows that, if the model performance is assessed in terms of Pearson correlation, RMSE, and MAD, the surges simulated with the ERA5 forcing fit better to the measured data (Fig. 3).*'

Lines 288-289: '*Similar results are obtained for MAD, which shows better performance for the simulations with ERA5 forcing at all locations. Only in Trieste does BC-UniGe achieve the same performance as BC-ERA5 for this metric.*'

Lines 319-320: '*Only in Caorle (Fig. 6.C) and Monfalcone (Fig. 6.E) does BC-ERA5 show better Pearson correlation, RMSE, and MAD; …*'

Lines 347-349: '*While simulations using ERA5 forcing generally show slightly better performance on traditional metrics such as RMSE, MAD, and Pearson correlation coefficient, …*'

Lines 356: '*Despite this, metrics such as Pearson correlation, RMSE, and MAD generally indicate better performance for ERA5 simulations.*'

Lines 402-406: '*Conversely, MAD, although it also experiences a form of "double penalty," reduces the impact of this effect compared to RMSE. As a result, the performance differences between simulations, noticeably above the 99th percentile, are generally more pronounced for MAD than for RMSE, better highlighting the superiority of BC-Unige. This enhanced differentiation is likely due to MAD's linear weighting of errors, which reduces the inflated impact of large deviations that characterizes RMSE.*'

Line 440: '*Traditional metrics like Pearson correlation, RMSE, and MAD favor a simulation (BT-ERA5) forced by a coarser…*'

Line 444: '*…indicators such as RMSE, MAD and Pearson correlation*.'

b) Have you also considered introducing RMSEp and RMSEc, either instead of or in addition to MADp and MADc? If so, why have you decided against using them?

R.b: Thank you for your valuable suggestion regarding the introduction of RMSE-based metrics (RMSEp and RMSEc) as alternatives or complements to the proposed MADp and MADc. Our decision to prioritize MAD-based metrics over RMSE-based ones was driven by the fact that RMSE inherently emphasizes larger errors due to its quadratic nature, making it more sensitive to outlier. While this can be beneficial in certain contexts, our primary focus was on providing a more balanced assessment of model performance across the entire distribution of observed data. The use of MAD, which is less sensitive to error outliers, allows us to better evaluate the general agreement between simulations and observations. This aspect is even more important in the case of phase error, as RMSE will

weight quadratically the model-observations differences, thus emphasizing even more than MAD the phase discrepancies.

4. Figures 3-7

a) When looking at Figure 7, it appears that BC-UniGe overestimates the extremes, while the other configurations underestimate them. A closer look at the models' abilities to simulate these extremes (99th percentile in Figure 6) actually shows that most metrics indicate BC-UniGe as the best model setup. When assessing the overall quality of the models, as highlighted in the other figures, it is actually challenging to identify a clear best model. It's true that the Pearson correlation coefficient and the RMSE indicate better performance for the ERA5 configurations, while MADp and MADc mostly identify BC-UniGe as superior. However, the differences in the metrics across the various configurations are often only of the order of millimetres. This should be highlighted and discussed in further detail. Maybe models of other regions (with higher variations in water levels) might even be better suited to demonstrate the benefits (and limitations) of your newly introduced metrics.

R.a: Thank you for this observation. You are correct that, when considering the total amount of data, the differences between the metrics for the simulations are often of the order of millimeters, with more pronounced differences in the surges above the 99th percentile, however, these figures are significant in relative terms. Analyzing the percentage variation of the proposed MADc metric using the total data, we find improvements with BC-UniGe ranging from 1.3% (Grado) to 9.3% (Trieste) compared to BT-ERA5, and 1.6% (Grado) to 10.3% (Trieste) compared to BC-ERA5. For surges above the 99th percentile, the improvements are even more pronounced, with BC-UniGe showing a 12% (Monfalcone) to 31.6% (Trieste) improvement over BT-ERA5, and a 4.1% (Caorle) to 20.2% (Trieste) improvement over BC-ERA5. This detailed analysis is now included in Section 4 of the revised manuscript, adding the following (lines 422-427): '*The differences between the simulation metrics are generally in the range of millimeters when considering the overall data, but these differences are significant in relative terms. For the MADc metric, BC-UniGe shows improvements ranging from 1.3% (Grado) to 9.3% (Trieste) compared to BT-ERA5, and from 1.6% (Grado) to 10.3% (Trieste) compared to BC-ERA5. The improvements are even more notable when focusing on values above the 99th percentile, where BC-UniGe outperforms BT-ERA5 by 12% (Monfalcone) to 31.6% (Trieste), and BC-ERA5 by 4.1% (Caorle) to 20.2% (Trieste).*'

b) Do you also have any insights into why model performance for the locations of Monfalcone and Caorle differs from that of the other locations?

R.b: Thank you for this question. The differences in results for Monfalcone and Caorle may be related to the locations of their tide gauges. The Caorle tide gauge is in a protected area inside the Livenza River, which is not fully represented by our model due to coastline resolution. Similarly, Monfalcone's tide gauge is positioned in front of a breakwater, which is also not fully represented. These local factors could affect the observed and simulated signals, particularly due to local effects around the tide gauges. This explanation has been added to the revised manuscript in section 4, adding the following (lines 431-435): '*As shown in Section 3, some discrepancies were observed in Caorle and Monfalcone, where BC-ERA5 achieved better performance in terms of MADc. A possible explanation for this could be related to the location of the tide gauges at these sites. The tide gauge at Caorle is situated in a protected area inside the Livenza River, a location not fully represented by the simulations due to the resolution of the coastline, even though a high-resolution model data was used. A similar issue is found in Monfalcone, where the tide gauge is located in front of a breakwater not fully represented by the coastline used in the model. These factors could affect the signals obtained from observations and simulations, primarily due to local effects at the tide gauge locations.*'

c) Furthermore, is it necessary to show 6 panels in all these figures (one for each tide gauge)? If most of the gauges show similar behaviour, it might be more effective to display only selected representative gauges and provide the rest in the supplementary material.

R.c: Thank you for your valuable feedback. We fully understand the importance of presenting information concisely. However, we believe that in this case including panels for all six locations is essential to provide a complete and clear understanding of the results without compromising clarity. To ensure readers gain a comprehensive insight, we have decided to retain all the panels in the main text.

5. LL343-345: "*In RMSE, "double penalty" is further amplified compared to MAD, as the penalizations due to the peak mismatch are squared. This means that phase errors have a disproportionately large impact on RMSE.*"

To provide a clearer picture, it would be beneficial to include the traditional MAD as an additional metric. Without this, it is challenging to determine whether the "double penalty" is the primary issue or if this amplification leads to the RMSE favouring the ERA5 configurations.

R: Thank you for this insightful comment. As mentioned earlier, we have now included the traditional MAD in the revised manuscript to provide a more comprehensive evaluation of the model performance. In our study, we chose to emphasize the MAD over the RMSE because it provides a more robust evaluation of model performance in the presence of outliers or extreme values, which can disproportionately influence RMSE due to its squaring of errors. This robustness is particularly important in storm surge simulations where occasional large deviations may occur.

Related to phase error, RMSE is particularly sensitive to this effect, which are common in time series data. If the timing of peaks and troughs in the simulation is slightly shifted relative to the observations. As we mention along the manuscript, RMSE can penalize the model twice: once for the difference in amplitude and again for the shift in time. This 'double penalty' effect can lead to an overestimation of the error, even when the overall shape and trend of the simulation closely match the observations. MAD is also affected by the double penalty, but this is somewhat mitigated by simply averaging the absolute differences, providing a more balanced assessment of the model's performance. Therefore, we believe that MAD better reflects the typical performance of our simulations.

6. LL371-372: "*This is because BC-UniGe is more prone to phase error than BT-ERA5, and is thus doubly penalized in indicators such as RMSE, MAD and Pearson correlation*."

Is this speculation or have you quantified it? It would be beneficial to understand the magnitude of phase differences between simulated and observed peaks across the different configurations of your model.

R: Thank you for this comment. We have now quantified the phase error between observations and simulations by identifying the peaks in the hourly time series and computing the model-observation phase error in hours. To do this, we identified the peaks in both the observations and the different simulations we carried out. Then, we estimated the difference, measured in hours, between the occurrence of the peak in the observations and the corresponding peak in the simulations. The mean values obtained show that the phase error in BC-UniGe is 3.1% higher than in BT-ERA5 and 4.5% higher than in BC-ERA5. This was mentioned in the revised version of the manuscript (lines 299-301): "*Additionally, these figures indicate that BC-UniGe simulations produce greater dispersion of data, likely due to a more frequent occurrence of phase error, which was quantified as 3.1% higher than in BT-ERA5, and 4.5% higher than in BC-ERA5*". The methodology followed to quantify the phase error was included on Section 2.3 of the revised manuscript, adding the following (lines 264-267): '*To quantify phase errors between observations and simulations, peaks in the hourly time series were identified for both observed and simulated data. The phase error was then calculated by measuring the time difference, in hours, between the occurrence of each peak in the observations and the corresponding peak in the simulations. This approach provided a direct assessment of the model's accuracy in capturing the timing of key events, such as storm surges.*'. Related to the results obtained, these were included in Section 3, lines 304-305.

<u>Technical corrections</u>

7. LL15-18: "*To illustrate this, we conducted 34-year high-resolution simulations for storm surge under barotropic (BT) and baroclinic (BC) configurations, using atmospheric data from ERA5 and a high-resolution downscaling of the Climate Forecast System Reanalysis (CFSR) developed by the University of Genoa (UniGe)*."

I have noticed this throughout your paper. Since your simulations deal with multiple extreme events, it would often be more fitting to use "storm surges" instead of "storm surge".

R: We appreciate your suggestion regarding the use of "storm surges" instead of "storm surge" throughout the paper. However, we would like to clarify that in this context, "storm surge" refers to the phenomenon as a whole rather than specific individual events. When we discuss storm surge in the manuscript, we are referring to the physical process that results in anomalous increases in sea level due to atmospheric conditions such as wind and pressure. Therefore, the term "storm surge" is used in a more general and conceptual sense rather than to describe discrete instances of the phenomenon.

8. LL88-90: "*The model has been already implemented in operational (Federico et al., 2017) and relocatable (Trotta et al., 2016) forecasting framework, and for storm surge events (Park et al., 2022; Alessandri et al., 2023)*."

The term "frameworks" should be used here instead of "framework".

R: Corrected.

9. LL104-105: "Finally, the conclusion show on Section 5 summarizes the key points of the study."

The word "show" should be omitted here.

R: Corrected.

10. LL112-114: "*ERA5 is relatively high resolution and accurate for a global reanalysis, although it is known to be affected by negative biases at high percentiles, particularly when is compared with measured wind speed (Pineau-Guillou et al., 2018; Vannucchi et al., 2021; Benetazzo et al., 2022; Gumuscu et al., 2023)*."

"[…], particularly when compared with measured wind speeds […]."

R: Corrected.

11. Figure 1

The red dashed box in panel (a) is barely visible; please use a thicker line. Additionally, the font sizes in panels (a) and (b) differ, and some fonts in panel (a) are extremely small. Have you also verified, whether the chosen colours in all your figures are readable for people with colour vision deficiencies? Please refer to the journal guidelines for further details.

R: Thank you for your comment. We have made improvements to Figure 1 by enlarging it and increasing the font sizes while respecting the font styles. It is important to mention that the differences in font size on the axes in panel (b) are intentional, as they indicate minutes in latitude and longitude. Additionally, we have ensured that the colors used are accessible to individuals with color deficiencies, in accordance with the journal's guidelines.

12. LL158-160: "*The observational data were acquired from Italian National Institute for Environmental Protection and Research (ISPRA), the Civil Protection of the Friuli-Venezia Giulia Region, and Raicich (2023)*."

"[…] the Italian National Institute […]"

R: Corrected.

13. L169: "*Both the model output and the observations were processed as follow to enable their intercomparability*."

"[…] as follows […]"

R: Corrected.

2 Referee #2

General reply: Thank you for your detailed review and helpful feedback. Your comments have been important in improving the clarity and overall quality of our manuscript. Below, we provide a point-by-point response to your remarks.

1. Introduction:

It is important to explicitly add the objective of the paper in the introduction. We infer it by the title and the discussion, but it is important to explicitly write it down. Also, it is important to introduce the problem before the objective (e.g. why there is a need to assess the error indicators for numerical models, even though they are established?).

R: Thank you very much for your suggestion. We have clarified the objectives of the work in the revised version of the manuscript, specifically in lines 68-70 of Section 1, adding the following: '… ***with two main objectives: first, to generate long-term databases of storm surge with a focus on accurately representing extreme values, and second, to analyze the ability of different metrics to capture the skill of the model.*****'**

2. Line 88: Fix citation.

R: Corrected.

3. Line 124: Fix citation

R: Corrected.

4. Line 131: Add the definition of nearshore used throughout the paper (it may differ across disciplines)

R: The term "nearshore" has been changed to "coastline," specifically in Section 2.2, line 133 of the revised manuscript.

5. Line 150-152: Consider entering the formula here instead.

R: The formulas has been added as Equations 1 and 2, lines 154-165 in the revised version of the manuscript.

6. Line 205: Could you also add the standard MAD for comparison, so to understand the effects of your changes on the direct metric it is based on? Also, since it is new metrics, it would benefit of some sort of "validation", or at least a sort of sensitivity analysis on the metrics based on well-defined synthetic time-series. Adding such comparison between your new metrics and the standard ones based on a synthetic (thus with specific known errors for testing) would enhance significantly this paper, especially considering the variability across the different metrics are small.

R: Thank you very much for your comment. The traditional MAD has been added to the revised version of the manuscript, along with its formula (Equation 7) and relevant discussion of its results. The added and/or modified lines in the revised version of the manuscript are the following:

Lines 230-234: ‘***Mean Absolute Deviation (MAD):***

$$\boldsymbol{MAD = \overline{|S - O|}} \tag{7}$$

A value closer to one indicates a better performance.’

Lines 283-284: ‘***The performance evaluation shows that, if the model performance is assessed in terms of Pearson correlation, RMSE, and MAD, the surges simulated with the ERA5 forcing fit better to the measured data (Fig. 3).***’

Lines 288-289: ‘***Similar results are obtained for MAD, which shows better performance for the simulations with ERA5 forcing at all locations. Only in Trieste does BC-UniGe achieve the same performance as BC-ERA5 for this metric.***’

Lines 319-320: ‘***Only in Caorle (Fig. 6.C) and Monfalcone (Fig. 6.E) does BC-ERA5 show better Pearson correlation, RMSE, and MAD; …***’

Lines 347-349: ‘***While simulations using ERA5 forcing generally show slightly better performance on traditional metrics such as RMSE, MAD, and Pearson correlation coefficient, …***’

Lines 356: ‘***Despite this, metrics such as Pearson correlation, RMSE, and MAD generally indicate better performance for ERA5 simulations.***’

Lines 402-406: ‘***Conversely, MAD, although it also experiences a form of “double penalty,” reduces the impact of this effect compared to RMSE. As a result, the performance differences between simulations, noticeably above the 99th percentile, are generally more pronounced for MAD than for RMSE, better highlighting the superiority of BC-Unige. This enhanced differentiation is likely due to MAD's linear weighting of errors, which reduces the inflated impact of large deviations that characterizes RMSE.***’

Line 440: ‘***Traditional metrics like Pearson correlation, RMSE, and MAD favor a simulation (BT-ERA5) forced by a coarser…***’

Line 444: '*…indicators such as RMSE, MAD and Pearson correlation*.'

Additionally, following your suggestion, MADp and MADc have been validated using sinusoidal time-series. For this purpose, we added the following in lines 269-275 of the revised manuscript: '*The proposed metrics were also validated using an idealized time series. A sinusoidal time series was generated to represent an observed parameter. Two simulated time series were then created: one with the same amplitude as the observation but shifted in time (introducing a phase error), and the other with the same phase as the observation but with half the amplitude. Various metrics were calculated and plotted on scatter plots (Fig. 1.S). The results indicated better performance for the simulation that underestimated the observations when assessed with Pearson correlation, RMSE, and MAD. In contrast, the time series that accurately captured the amplitude was penalized for the phase error, which negatively affected its performance on these metrics. However, the proposed MADp and MADc metrics identified it as the better model.*'. The figures for these results are available as supplementary material in the revised version of the manuscript.

7. Results: Consider adding a small section where you discuss the storm patterns based on the dataset (e.g. histograms, frequency, so forth). It does not need to be long since the core of the paper is about testing the skill assessment methods. Nonetheless, it is important for the reader to conceptualize with the sort of storm patterns seen here.

R: Thank you very much for the suggestion. A brief description was added in Section 3, lines 267-281 of the revised version of the article, adding the following: '*The Probability Distribution Estimates (PDE) and Empirical Cumulative Distribution Functions (ECDF), available in Fig. 2S to 7S, show that BC-UniGe better represents the higher values of storm surge when compared with observations, particularly when considering values above the 99th percentile However, some overestimations are noticeable in Caorle and Monfalcone with BC-UniGe. In contrast, simulations with ERA5 forcing tend to underestimate these higher values, which is more noticeable for BT-ERA5.*'. Figures of the described results will be available in the supplementary material.

8. Radar charts: Add on the caption of the applicable figures throughout the manuscript that values on the fringe refer to better performance.

R: This information has been added in the revised version of the manuscript.

3 Referee #3

General reply: Thank you for your thorough review and valuable comments. We sincerely appreciate the time and effort you have taken to provide feedback on our work. We have carefully considered your suggestions and made corresponding revisions. Below, we provide detailed responses to each of your comments.

1. In the article, the authors discuss several modelling choices that could potentially impact the results. However, there is no explanation for the selection of SHYFEM for this modelling effort. I suggest that the authors elaborate on their choice, discussing the advantages and disadvantages of using SHYFEM.

R: Thank you for the comment. We agree that providing a rationale for our choice of SHYFEM is important for the context of our study. We have now included the reasons for selecting SHYFEM, which is related to the works already carried out with the model (references mentioned in lines 89-90). This information has been added to Section 1 of the revised manuscript, specifically on lines 90-93, adding the following: '*The choice of SHYFEM is driven by its flexibility in handling complex bathymetry and irregular coastlines through its unstructured-grid framework, allowing for higher resolution in critical areas. Additionally, its successful implementation in operational and relocatable forecasting frameworks, and storm surge events, confirms its reliability for this study*'.

2. The comparison between BT-ERA5, BC-ERA5, and BC-UniGe is influenced by the use of different atmospheric simulations. For the first two, the base reanalysis model is ERA5, whereas the comparison is later shifted to a downscaled CFSR product. Given the sensitivity of extreme results for these two models, this comparison seems somewhat unfair. While I understand the role of metrics and agree with the overall message of the paper, you are comparing different products. Although ERA5 and CFSR might be similar for average statistics, it is unclear how they compare for extremes, particularly in an enclosed basin like the Adriatic Sea. The paper should clearly state this, or if possible, show the performance of the native ERA5 and CFSR storm surge products in the region. Additionally, consider discussing what a WRF ERA5 downscaling would contribute to the analysis.

R: We thank the reviewer for this comment. We are fully aware of the differences between ERA5 and a downscaled CFSR. We mention explicitly these differences in the article, for example when we state that (lines 119-120):
"*Since ERA5 is relatively coarse for local studies and exhibits significant underestimation of extremes, we employed an alternative approach using a high-resolution (3.3 km) atmospheric downscaling developed by the University of Genoa (UniGe)*."

And also (lines 365-368):
"*From an epistemic point of view BC-UniGe is a significantly more sophisticated model compared to BT-ERA5. Not only does it employ a higher resolution forcing. It also takes into account the baroclinicity and the vertical motion within the*

***water column, whereas the barotropic configuration of BT-ERA5 approximates the ocean as a 2D sheet only subject to vertically uniform motions and waves*".**

The focus of this publication is on the fact that we know BC-UniGe is better, but this does not necessarily translate into better values of error indicators such as RMSE. We acknowledge that a WRF-ERA5 downscaling could improve the simulations, especially for higher surge values where discrepancies were observed. However, as the primary objective of our study is to analyze the impact of different statistical metrics on the evaluation of numerical simulations, we decided to focus on this aspect and not delve deeply into the atmospheric forcings.

3. In comparing the tide gauge time series with the different simulations, there are noticeable differences between the modelled and observed time series. I find puzzling that the impact of waves is not mentioned in the paper. What effect would waves have on these comparisons? I suggest the authors include comments on whether waves play a crucial role in the storm surge levels in the Adriatic Sea.

R: Thank you for this observation, which is indeed highly relevant. We recognize that waves can significantly influence storm surge levels. However, as mentioned in response to your previous comment, the main focus of this study is to demonstrate how different statistical metrics can affect the evaluation of numerical simulations. While the role of waves is undeniably important, we have chosen to limit the scope of our discussion in this paper to the metrics themselves. We greatly appreciate your insightful comments, and we will consider exploring the impact of waves in future studies.