

General comments:

The manuscript **“MAROR: a multipoint autoregressive network for forcing-driven, uncertainty-calibrated reconstruction of hourly coastal sea level from 1940 to the present”** applied a deep learning model to reconstruct water levels time series of tidal gauges. The authors further extended the capabilities of HIDRA3 model, allowing it to reconstruct historical records, which is the innovation on their research.

Although the manuscript is well-written and contributes to advancing the knowledge about deep learning models' applications to surge modelling, I believe it is still not suitable for publication due to some points that needs further attention. Based on that, I recommend a major review so the manuscript can be accepted.

Please, find my comments below.

Specific comments:

#1 - The methodology is clear and well designed, adequate for the purpose of the manuscript. However, further explanation is necessary to facilitate the reproduction of the results.

Methodology

The forcing have a 1 ° resolution. However, most storms have a radius around several kilometres, which is much lower than the forcings resolution. How the model would behave when applied to predict the storm surge of such events?

Highlighting this limitation and connecting this with the scale of the physical process that act during a surge could contribute to better understanding the relevance of each process to the total surge. In case the model presents satisfactory performance in the peak prediction of low max radius events, this could imply that the low-resolution data have signals that allow the model to correctly predict the surge peaks.

#2 - Section 2.4.2 – The authors used a 72 hour window for the inputs to predict the next hour. However, each component takes different time to disturb the water level in an observation point. For example, tides are more regular and predictable, so 72 hours may be sufficient, depending on the tidal constituents, but atmospheric forcings are subject to a series of other factors. In the case of a

storm with a slow translational speed and strong winds, it may take more than 72 hours for the storm surge to reach the coast. Have the authors tested different intervals? It would be interesting to see the model sensitivity to different time windows.

3 - Section 2.5.3 – As the authors are predicting a time series, adding the Kling Gupta efficiency to the error metrics would allow the authors to assess not only the bias and the correlation (which are already considered), but also variability of the model predictions.

Reference: <https://doi.org/10.1016/j.jhydrol.2009.08.003>.

4 -Section 3.1 – Adding a workflow showing all the process that were applied to the raw data would help the reader to better visualise and understand all the steps required to generate the train, test and validation data.

5 - Section 3.4 – In this section, adding a time series plot of the annual maximum surge in each data set (train, test and validation) would be useful to visualize if the extremes at each set are equivalent. This plot would allow the readers to have a better understanding of the data that feeds the model, and the magnitude of the output variable.

Additionally, it allows the readers to better evaluate the model performance, in case where the train data presents different trends and extreme values than the test data. For example, if the peak of the validation/test dataset is much lower than then peak of the training dataset, comparing the peak performance of the model only in these datasets would be misleading. I suggest that the authors could select at least one gauge in each region to demonstrate it.

#6 - Section 4.3:

Line 512: The authors should add a reference to support the following statement, or refer to the data they are using that leads to this conclusion.

“It is worth noting that tropical cyclones drive only a minority of the strongest surges in these basins, about 13% of the largest events in the Gulf of Mexico and 31% in southern Japan, with extratropical storms accounting for the rest, yet tropical cyclones are the most extreme and most localised events.”

7 - The authors acknowledge that ERA5 tends to underestimate hurricane-driven surges due to its resolution constraints (Line 529 onwards). However, the authors do not provide enough evidence to state that the only reason for the underestimation is due to ERA5 resolution and not introduced by other inputs or the network architecture, even though I agree that their hypothesis is plausible.

To support their statement, I would recommend to bias correct the ERA5 inputs to eliminate this source of uncertainty, even though the problem is not necessarily with the magnitude, but with the resolution itself. This would ensure that the bias of their model is not coming from ERA5, but due to something else (e. g. model architecture or other inputs), in case a bias persists.

Technical corrections:

Reference: The reference on line 742 (Klambauer et al, 2017) has a broken link that leads to a non existing page.