Reviewer Comments

The manuscript presents MAROR, a multipoint autoregressive neural network for reconstructing hourly coastal sea level from 1940 to the present using atmospheric forcing and, where available, satellite altimetry. The topic is relevant for coastal flood-hazard assessment, particularly because the manuscript combines long-term reconstruction, predictive uncertainty estimation, and evaluation across 373 tide gauges in different coastal settings. However, in its present form, the manuscript requires major revision. The main concerns are the independence of the validation procedure, the potential effects of data preprocessing on genuine extreme events, insufficient evidence for long-term uncertainty calibration, inconsistencies in the comparisons with existing models, and the treatment of reconstruction uncertainty in extreme-value analysis. The claims of physical consistency and the use of model outputs for observational quality control also require stronger supporting evidence and more cautious wording.

Comment #1: The Introduction and Discussion cover existing sea-level modeling approaches, but the literature review would benefit from a broader discussion of long-term storm-surge reconstruction, the representation of extreme events in data-driven models, and the implications of sea-level change for coastal inundation and exposure. These aspects would help clarify both MAROR's methodological contribution and its potential applications in coastal hazard assessment. The authors should distinguish advances in reconstructing hourly variability and extremes from studies assessing the longer-term consequences of rising sea levels. The following studies provide examples covering these aspects and may be considered where appropriate; they are not intended as a restrictive reference list:

- Tadesse, M. G., & Wahl, T. (2021). A database of global storm surge reconstructions. Scientific Data, 8, Article 125. https://doi.org/10.1038/s41597-021-00906-x
- Yang, L., Jin, T., Xiao, M., Gao, X., Jiang, W., & Li, J. (2023). Extreme events and probability analysis along the United States east coast based on high spatial-coverage reconstructed storm surges. Geophysical Research Letters, 50, Article e2023GL103492. https://doi.org/10.1029/2023GL103492
- Campos-Caba, R., Camus, P., Mazzino, A., Vousdoukas, M., Tondello, M., Federico, I., Causio, S., & Mentaschi, L. (2026). Storm surge dynamics in the northern Adriatic Sea: Comparing AI emulators with high-resolution numerical simulations. Natural Hazards and Earth System Sciences, 26, 3683–3706. https://doi.org/10.5194/nhess-26-3683-2026
- Zengin, E. (2023a). A combined assessment of sea level rise (SLR) effect on Antalya Gulf (Türkiye) and future predictions on land loss. Journal of the Indian Society of Remote Sensing, 51, 1121–1133. https://doi.org/10.1007/s12524-023-01694-0
- Zengin, E. (2023b). Inundation risk assessment of Eastern Mediterranean coastal archaeological and historical sites of Türkiye and Greece. Environmental Monitoring and Assessment, 195, Article 968. https://doi.org/10.1007/s10661-023-11549-3

Comment #2: Section 3.1 applies a five-standard-deviation outlier check to the storm-surge residual. This procedure may flag genuine extreme events. Please clarify how flagged values were assessed and demonstrate that the cleaning procedure preserves documented storm peaks and does not bias the extreme-value analysis.

Comment #3: Please clarify whether normalization parameters, trends, and tidal constituents were estimated using training data only or the full record. Also, explain how observation-based input windows crossing training/test boundaries were handled to prevent information leakage.

Comment #4: The claim of calibrated uncertainty requires quantitative evidence from independent test years. Please report the coverage and average width of the nominal 90% prediction intervals, including during extreme events, and assess whether calibration remains reliable over multidecadal reconstructions.

Comment #5: Some comparisons with the previous deep-learning model overlap MAROR's training years and use different observational references. Please base the main performance comparison on common held-out periods and identical observations, and revise the claims of superiority accordingly.

Comment #6: A small overall bias does not demonstrate the absence of long-term drift. Please compare the extended reconstructions before and after the scheduled sampling and show whether the bias remains stable across successive decades.

Comment #7: Setting SLA inputs to zero before 1993 does not guarantee equivalence to a separately trained no-SLA model. Please assess whether the introduction of altimetry creates discontinuities in reconstruction statistics, particularly in southern Japan.

Comment #8: Please clarify whether the GEV confidence intervals incorporate reconstruction uncertainty or only uncertainty in fitting the distribution. Also assess sensitivity to the 50% annual data-coverage threshold, as incomplete years may miss the true annual maximum.

Comment #9: The tropical-cyclone analysis allows reconstructed peaks within ±24 hours of observed peaks, potentially masking timing errors. Please report peak timing errors separately and distinguish performance on training and test events.

Comment #10: Describing MAROR as physically plausible "by construction" and an independent reference for quality control requires stronger justification. Please validate the proposed detector against documented or synthetic clock errors and assess whether genuine local sea-level variability could be incorrectly flagged.