We thank the reviewer for the careful and constructive evaluation of our manuscript. We greatly appreciate the insightful comments, which have helped us improve the clarity, methodological transparency, and contextualization of our work. Below we provide a detailed, point-by-point response to each comment. In the remainder, reviewer comments are reported in bold-italics, followed by our responses. All changes are incorporated in the revised manuscript, with explicit references to sections, figures, and supplementary material.

During the preparation of the revised manuscript, we identified an implementation issue in the inference pipeline: predictor variables in the testing and case-study datasets were inadvertently standardized using statistics computed from the testing set itself rather than the scaler fitted on the training data. This was inconsistent with the preprocessing applied during model training. The pipeline has now been corrected so that all predictors in the testing set and in the November 2022 inference experiment are standardized using the scaler derived from the training data.

We are also grateful to the reviewer for several comments that prompted us to re-examine and further improve the implementation of the recurrent architectures. During the revision process, we refined the recurrent models by fully adopting a sequence-to-sequence formulation combined with Truncated Backpropagation Through Time (TBPTT), thereby improving the consistency with which temporal dependencies are represented during both training and inference. The revised manuscript now includes a detailed description of this implementation in Section 2.4. In addition, several reviewer comments motivated new sensitivity analyses and methodological clarifications, which have strengthened both the robustness and transparency of the study.

After implementing these corrections, we recomputed the full set of experiments, including the 40 runs for each emulator architecture and the November 2022 case study. The figures and quantitative results have been updated accordingly. While quantitative differences are observed, the overall performance patterns and qualitative behavior of the models remain consistent with the original submission, and the main conclusions of the manuscript are unchanged.

Overall, these revisions improve the methodological consistency of the study while preserving the central findings, namely that the interaction between loss-function design and model architecture plays a central role in accurately reproducing extreme storm surge events.

General Comments

Significant efforts are made to use best practices for machine learning and to evaluate the statistical significance of the results. This is the primary strength of the paper. Additionally, the application of ML for storm surge to the Adriatic Sea is relatively novel. The use of multiple training metrics and separate consideration of extreme surges is also a plus. The paper does not clearly explain the ML problem formulation. Additionally, there is little explanation of why

storm surge prediction is important, and no clear indication of how the developed models would be used.

We thank the reviewer for recognizing the strengths of the methodological design and statistical evaluation. In response to the concerns raised, we have extended the description of the ML problem formulation across Section 2.

We have also expanded the Introduction (Section 1) to better contextualize the societal and economic relevance of storm surge in the northern Adriatic Sea and clarified the intended operational role of the developed models within a statistical downscaling framework. The following can now be found in the revised manuscript (L41-50): *The northern Adriatic Sea represents a vulnerable coastal environment in Europe. Its shallow bathymetry, funnel-shaped geometry, and frequent intense sirocco wind events make the Venice lagoon and surrounding low-lying coastal areas highly susceptible to flooding. Several extreme events have demonstrated the societal and economic consequences of Adriatic storm surges, including the catastrophic flood of 4 November 1966, which submerged most of Venice and caused extensive damage to infrastructure and cultural heritage (De Zolt et al., 2006), as well as more recent events in 2018 and 2019 that led to widespread urban flooding and economic losses of hundreds of millions of euros (Ferrarin et al., 2020; Umgiesser et al., 2021). The activation of the MoSE flood barrier system during the November 2022 surge further underscores both the persistent hazard and the operational importance of accurate surge prediction (Mel et al., 2023). Although Adriatic storm surges are generally smaller in magnitude than those associated with tropical cyclones, their interaction with densely populated and culturally significant coastal zones makes reliable surge modeling essential for risk management and infrastructure planning.*

Specific Comments

1. The introduction talks about storm surge... but then the paper focuses on predicting water levels at tidal gauges. Storm surge specifically refers to the excess water height above the regular tide caused by an extreme weather event. This distinction needs to be made clear in the paper.

We thank the reviewer for highlighting this important conceptual distinction. Section 2.2 (Predictand) has been substantially expanded to explicitly describe the storm surge estimation procedure applied to both observations and SHYFEM-MPI output. We now clearly define storm surge as the filtered non-tidal residual obtained after detrending, yearly harmonic analysis using T-Tide, and low-pass filtering with a 13-hour cutoff. The manuscript consistently distinguishes between total sea level and storm surge throughout. The following can now be found in the revised manuscript (L108-119): *Observed data from tide gauges were used as the predictand. In this study, the target variable corresponds to storm surge, defined as the meteorologically driven, non-tidal residual component of sea level. The storm surge estimation follows the procedure described in Campos-Caba et al. (2024) and consists of three main steps. First, the raw sea-level time series were centered to zero mean and linearly detrended to remove long-term changes. Second, harmonic analysis was performed separately for each*

calendar year using the T-Tide MATLAB package (Pawlowicz et al., 2022). The non-tidal residual (NTR) was computed as the arithmetic difference between the total sea level and the reconstructed tidal signal.

Finally, to isolate the pure storm surge signal (hereafter also referred to as "surge"), a low-pass filter was applied to the non-tidal residual following Park et al. (2022). A cut-off period of 13 hours was adopted, consistent with the mixed semidiurnal tidal regime of the northern Adriatic Sea (Lionello et al., 2021). The resulting filtered non-tidal residual constitutes the predictand used for training and evaluation of the ML emulators. To ensure strict intercomparability, the identical preprocessing chain was applied to the SHYFEM-MPI output prior to performance assessment.

2. Traditionally, storm surge forecasting focuses on regions subject to tropical cyclones, such as the North Atlantic, the Indian Ocean, or the Pacific Ocean. They are responsible for significant loss of life and property damage, which justifies significant forecasting efforts. What is the impact of storm surge in the Adriatic Sea specifically?

We appreciate this important request for clarification. While Adriatic storm surges are typically smaller in magnitude than those associated with tropical cyclones, the northern Adriatic Sea is one of the most vulnerable coastal environments in Europe due to its shallow bathymetry, funnel-shaped geometry, and the exposure of densely populated and culturally significant areas such as Venice.

Several extreme events have demonstrated the substantial societal and economic impacts of Adriatic storm surges, including the catastrophic 1966 flood and more recent events in 2018 and 2019 that caused widespread flooding and major economic losses. The operation of the MoSE flood barrier system during the November 2022 event further underscores the operational importance of accurate surge prediction in this region.

We have expanded the Introduction (Section 1) to explicitly describe the historical, societal, and infrastructural relevance of storm surge in the northern Adriatic Sea, thereby strengthening the regional motivation of the study. The following was added on L41-50: *The northern Adriatic Sea represents a vulnerable coastal environment in Europe. Its shallow bathymetry, funnel-shaped geometry, and frequent intense sirocco wind events make the Venice lagoon and surrounding low-lying coastal areas highly susceptible to flooding. Several extreme events have demonstrated the societal and economic consequences of Adriatic storm surges, including the catastrophic flood of 4 November 1966, which submerged most of Venice and caused extensive damage to infrastructure and cultural heritage (De Zolt et al., 2006), as well as more recent events in 2018 and 2019 that led to widespread urban flooding and economic losses of hundreds of millions of euros (Ferrarin et al., 2020; Umgiesser et al., 2021). The activation of the MoSE flood barrier system during the November 2022 surge further underscores both the persistent hazard and the operational importance of accurate surge prediction (Mel et al., 2023). Although Adriatic storm surges are generally smaller in magnitude than those associated*

with tropical cyclones, their interaction with densely populated and culturally significant coastal zones makes reliable surge modeling essential for risk management and infrastructure planning.

3. To improve reproducibility, the full list of features should be enumerated in a table.

We thank the reviewer for this constructive suggestion. To improve transparency and reproducibility, we have added a new table (Table 1) in Section 2.3 that explicitly enumerates all predictor variables used in the ML emulators, including their data sources, preprocessing steps (e.g., PCA dimensionality reduction), and resulting feature dimensionality per location. This structured summary complements the textual description and ensures that the feature construction process is fully documented.

4. It appears that separate models are trained for each prediction location, but that the predictor inputs to these models are identical. If so, have you considered including location specific predictors? Also, this should be clearly explained.

We thank the reviewer for this observation and the opportunity to clarify. Separate ML emulators were trained independently for Punta della Salute and Trieste. Although the same classes of basin-scale predictors (SSH, mean sea level pressure, wind stress components, and tides) were considered, predictor extraction and PCA dimensionality reduction were performed separately for each location using spatial domains centered around the respective tide gauges. Consequently, the predictor matrices are location-specific and reflect differences in spatial footprint and local coastal dynamics. We have clarified this explicitly in the following sections of the revised manuscript:

- Sections 2.3, L164-168: *Predictor extraction and PCA were performed separately for each location using the corresponding spatial domain (Figure 1a). Although the same types of basin-scale variables (sea surface height, mean sea level pressure, wind stress components, and tides) were used, the resulting predictor matrices are location-specific, reflecting differences in spatial footprint and local coastal dynamics. For reproducibility and clarity, Table 1 summarizes the predictor variables used in the ML emulators, their data sources, preprocessing steps, and resulting feature dimensionality.*
- Section 2.5, L324-328: *Considering the different emulator configurations and loss functions, a total of 12 ML emulators were implemented (Table 3). It is important to note that separate ML emulators were trained independently for Punta della Salute and Trieste. Although the same classes of basin-scale predictors were considered, predictor extraction, PCA dimensionality reduction, and model training were performed separately for each site. This ensures that each emulator learns a location-specific statistical mapping between large-scale forcing and local surge response.*

5. The most important predictor is sea surface height, which is an output from a physics-based ocean circulation model. This means that in any application context, the physics-based model would need to be run first. Consequently, the models in this work function as correctors to an existing physics-driven model of water height. This can result in greater accuracy than the original model but not increased computational efficiency. The cost of running Med-MFC needs to be added to the ML model training/inference cost for purposes of computational performance comparison.

We thank the reviewer for this important clarification. Our approach is intentionally framed as a statistical downscaling paradigm rather than a standalone replacement of basin-scale ocean circulation models. In operational coastal forecasting chains, basin-scale models such as Med-MFC routinely provide background SSH fields, which are subsequently refined by higher-resolution regional models.

In this study, the ML emulators replace the high-resolution coastal refinement step (SHYFEM-MPI), not the basin-scale Med-MFC simulation itself. The computational comparison therefore refers specifically to substituting the regional dynamical downscaling component with a statistical refinement model. We have clarified this distinction in the following sections of the revised manuscript:

- Section 2.3 (Predictors), L140-148: *The inclusion of basin-scale sea surface height (SSH) from the Med-MFC reanalysis reflects a deliberate choice aligned with a statistical downscaling framework, rather than an attempt to replace physics-based models. In coastal downscaling, large-scale ocean models routinely provide SSH boundary conditions that are subsequently refined by higher-resolution regional or coastal models. In this study, the ML emulators are designed to perform an analogous task: they take available basin-scale SSH fields together with atmospheric predictors and statistically refine them to tide gauge-scale storm surge signals. In this sense, the ML emulators function as data-driven coastal correctors of basin-scale output rather than standalone substitutes for ocean circulation models. This approach differs from purely atmospheric-driven time-series forecasting and instead mirrors established dynamical downscaling chains, focusing ML capacity on resolving coastal-scale processes that coarse models cannot explicitly capture.*
- Section 4 (Discussion), L612-620: *An important methodological choice in this study is the use of basin-scale sea surface height from Med-MFC as a predictor for the ML emulators. While several data-driven storm surge studies rely exclusively on atmospheric forcing, our approach is intentionally framed as statistical downscaling rather than standalone time-series downscaling. By leveraging basin-scale SSH, freely and consistently available through Copernicus services, the emulator refines the large-scale ocean signal to the local coastal response, in direct analogy with the role played by high-resolution dynamical models in traditional downscaling chains. This design choice is not circular, as the emulator does not attempt to reproduce the basin-scale solution, but instead learns the systematic transformations required to resolve local surge dynamics that are unresolved at coarse resolution. At the same time, this choice implies that emulator*

performance is conditional on the availability and quality of the parent ocean model, a limitation that we explicitly acknowledge.

- Section 4 (Discussion), L780-788: *Importantly, the SSH predictor used for training (Med-MFC reanalysis) can be replaced in forecasting applications by operational Copernicus products such as the Sea Physics Analysis and Forecast system (Med-Physics). The November 2022 case study demonstrates this compatibility: the trained emulators were applied using Med-Physics SSH fields, without requiring retraining or reconfiguration. This indicates that the framework is directly transferable to forecast scenarios in which basin-scale SSH is provided by an operational forecasting system. In ensemble forecasting contexts, the emulator would ingest ensemble members of basin-scale SSH and atmospheric forcing fields. The computational cost associated with generating such ensembles remains tied to the basin-scale ocean model. However, the emulator replaces the expensive coastal refinement step for each ensemble member. Consequently, the efficiency gains arise from eliminating repeated high-resolution coastal simulations rather than from replacing the basin-scale circulation model itself.*
- Section 4 (Discussion), L790-792: *In terms of computational efficiency, the contrast between dynamical downscaling and the ML approach is striking. The computational comparison discussed below concerns the replacement of the high-resolution coastal downscaling model (SHYFEM-MPI), while treating basin-scale Med-MFC output as an external operational input.*
- Section 4 (Discussion), L803-809: *It is important to emphasize that the computational comparison presented in this study refers to replacing the high-resolution regional downscaling model (SHYFEM-MPI), not the basin-scale Med-MFC simulation itself. Med-MFC is treated here as an operationally available large-scale input, analogous to boundary conditions in traditional dynamical downscaling systems. The efficiency gains demonstrated by the ML emulators therefore arise from replacing expensive coastal-scale dynamical refinement rather than eliminating the need for basin-scale ocean simulations. We acknowledge that in ensemble forecasting applications, multiple realizations of basin-scale forcing would still be required. In this context, the ML emulators would act as lightweight refinements of ensemble SSH fields rather than full substitutes for circulation models.*

6. When discussing the potential applications of these models, it needs to be made clear that they rely on the sea surface output of Med-MFC, which is the same physical quantity as what the models intend to predict (sea surface height). Using them in an ensemble forecast scenario will require multiple evaluations of Med-MFC. It may be worthwhile to train a model that relies only on wind and pressure fields, which is more suitable for integration into forecasting applications and has the expected performance advantages for a machine learning model (compared to physics-based ocean circulation models).

We thank the reviewer for this insightful comment. As mentioned in the response to the previous comment, we have clarified in the revised manuscript that the ML emulators are framed as

statistical coastal refinements of basin-scale sea surface height fields rather than standalone substitutes for ocean circulation models. The SSH predictor used for training (Med-MFC reanalysis) can be replaced in operational settings by basin-scale forecast products such as the Copernicus Sea Physics Analysis and Forecast system (Med-Physics).

Indeed, in the November 2022 case study, the trained emulators were successfully applied using Med-Physics forecast SSH fields, demonstrating compatibility with operational products without retraining. This transferability is now explicitly discussed on Section 4, L780-788, which reads as follows: *Importantly, the SSH predictor used for training (Med-MFC reanalysis) can be replaced in forecasting applications by operational Copernicus products such as the Sea Physics Analysis and Forecast system (Med-Physics). The November 2022 case study demonstrates this compatibility: the trained emulators were applied using Med-Physics SSH fields, without requiring retraining or reconfiguration. This indicates that the framework is directly transferable to forecast scenarios in which basin-scale SSH is provided by an operational forecasting system. In ensemble forecasting contexts, the emulator would ingest ensemble members of basin-scale SSH and atmospheric forcing fields. The computational cost associated with generating such ensembles remains tied to the basin-scale ocean model. However, the emulator replaces the expensive coastal refinement step for each ensemble member. Consequently, the efficiency gains arise from eliminating repeated high-resolution coastal simulations rather than from replacing the basin-scale circulation model itself.*

In response to the reviewer's suggestion, we conducted additional experiments using only atmospheric predictors (mean sea level pressure and wind stress components), excluding basin-scale SSH. These atmospheric-only configurations represent a fully standalone forecasting-style emulator. As expected, performance decreases moderately relative to SSH-informed models. However, the results confirm that purely atmospheric-driven models remain viable alternatives when basin-scale ocean forecasts are unavailable. This complementary configuration is now discussed in the revised manuscript Section 4, L622-633, which reads as follows: *To further assess the operational independence of the proposed framework, additional experiments were conducted using only atmospheric predictors (mean sea level pressure and wind stress components), excluding basin-scale SSH and tide, while maintaining the same configurations described in Sections 2.4 and 2.6. These atmospheric-only configurations represent fully standalone forecasting-style emulators. As expected, overall performance metrics degrade relative to SSH-informed models, reflecting the loss of explicit basin-scale ocean state information. Pearson correlation and slope values decrease moderately, and error metrics increase across both locations (Figure S9). However, the degradation is not drastic, and the atmospheric-only emulators retain substantial skill in reproducing surge variability. Notably, the percentile structure and the scaling of higher surge values remain reasonably well captured, indicating that atmospheric forcing alone contains significant predictive information for extreme events. These results suggest that while basin-scale SSH provides valuable*

additional skill, particularly for overall variance reproduction, purely atmospheric-driven emulators remain viable alternatives in contexts where ocean model output is unavailable. The comparison highlights a clear trade-off between physical completeness and operational independence.