GENERAL COMMENTS

This manuscript presents a hybrid machine-learning (ML) framework for short-term storm-surge nowcasting in the southern North Sea. Rather than relying exclusively on observations or numerical modelling, the authors combine three complementary sources of information: tide-gauge observations from 22 stations, ERA5 atmospheric forcing, and storm-surge simulations from the operational COHERENS hydrodynamic model. The implemented ML emulators predict surge residuals at 10-minute intervals using a 12-hour input history and initially generate forecasts for the subsequent two hours. Longer forecast horizons, extending up to 12 hours, are obtained recursively.

The results indicate that combining numerical-model outputs with observations can provide more accurate and operationally useful nowcasts than using either source of information independently. Overall, the study lies at the intersection of operational coastal forecasting, hybrid modelling, scientific machine learning, and coastal-hazard prediction.

The manuscript addresses an important operational problem through an interesting hybrid modelling strategy. The work is generally well written, technically sound, and potentially suitable for publication. However, in its current form, several methodological choices would benefit from further justification, and the scientific contribution could be demonstrated more convincingly.

In my view, the principal limitation is not the ML methodology itself, but rather the relatively limited scientific analysis of why the hybrid framework improves the predictions, under which conditions these improvements occur, and to what extent the proposed approach can be expected to generalize beyond the present application. Addressing these aspects, particularly in the Discussion, would substantially strengthen the manuscript.

SPECIFIC COMMENTS

L03–04: "*while purely data-driven models suffer from the limited duration of tide-gauge records*."

This statement may give the impression that data-driven models can only be trained using observations. I suggest slightly reframing the problem statement to acknowledge that outputs from numerical models can also provide training targets for ML emulators, particularly where long observational records are unavailable.

L10–11: "*… and modestly better than the pure data-driven baseline (0.061 m; 0.043–0.100 m)*."

The meaning of "pure data-driven baseline" was not immediately clear to me. Does this refer to an emulator trained using only observations, or to a model that excludes COHERENS information? I suggest defining this terminology explicitly when it is first introduced.

L13: "*the COHERENS baseline*."

Here, COHERENS is described as the baseline, whereas the preceding lines also refer to a "pure data-driven baseline." Please consider distinguishing these baselines more explicitly and consistently throughout the manuscript.

L14–15: "*The approach should generalize to other coastal regions with comparable observational coverage*."

This statement appears stronger than can be demonstrated from the experiments presented. The study investigates a single geographical region, and no transfer-learning, cross-region, or leave-one-station-out experiments are presented. I therefore suggest softening this conclusion or clearly presenting it as a hypothesis for future investigation rather than a demonstrated property of the framework.

L38–39: "*This article investigates how data-driven models and artificial intelligence can improve this final, yet crucial, stage of the decision-making process*."

This sentence appears to express an important objective of the study. I wonder whether it would be more effective if positioned later in the Introduction, together with the specific objectives and contributions of the manuscript.

L43–50: Several references concerning ML approaches for storm-surge modelling are introduced here. However, the discussion does not sufficiently cover previous data-driven and machine-learning storm-surge applications in the North Sea or closely related areas. This makes it difficult to fully assess the regional and methodological novelty of the present work. I encourage the authors to consider relevant studies such as:

- Krieger et al. (2025): https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2024GL111558
- Schaffer et al. (2025): https://doi.org/10.5194/nhess-25-2081-2025
- Royston et al. (2013): https://www.sciencedirect.com/science/article/pii/S0165011412004289
- Kristensen et al. (2026): https://arxiv.org/html/2601.02090

Including these and other relevant regional studies would help place the proposed methodology within the existing North Sea literature and clarify more precisely how the present contribution differs from previous work.

L51–52: "*Neural networks typically require large training datasets to learn complex feature relationships, which is difficult when observations are scarce or only short historical records are available*."

This is a reasonable point. However, I suggest also acknowledging that numerical-model simulations can provide training targets for ML emulators when sufficiently long observational records are unavailable. Observations are generally preferable when the objective is to reproduce measured conditions, since an emulator trained exclusively on numerical-model output may also learn systematic biases present in the parent numerical model. Nevertheless, numerical simulations can provide valuable training data where observational coverage is limited.

L59–60: "*In this paper, rather than replacing the physics with a machine-learning surrogate, we exploit the physics at the data level*."

Does "the physics" refer specifically to information provided by the physics-based COHERENS model? If so, I suggest stating this more explicitly. The current wording is somewhat abstract and could be interpreted in different ways.

L64–65: "(iii) *we carry out a systematic comparison against both the operational COHERENS hydrodynamic baseline and a pure data-driven LSTM+Transformer baseline*."

This would be an appropriate place to explicitly define what is meant by "pure data-driven," particularly which predictors are included or excluded relative to the hybrid configuration.

Introduction: More generally, I suggest revisiting the flow of the final three paragraphs. A clearer progression from (i) existing approaches and limitations, to (ii) the research gap, and finally to (iii) the objectives and specific contributions of the present study would strengthen the motivation.

L84–85: "*flagged values were discarded and replaced by linear interpolation when the gap was shorter than 2 hours (12 observations)*."

Could the authors provide additional information on the extent of this preprocessing? For example, it would be useful to report the percentage of observations that were interpolated, the amount of missing data at each station, and whether interpolation affected periods containing extreme surge events. This information would help assess the potential influence of preprocessing on model training and evaluation.

Fig. 1: I understand that the M2 cotidal and cophase information is shown because of the importance of tidal dynamics in the study area. However, its relevance to the subsequent analysis is not entirely clear, particularly because it does not appear to be discussed further. It may be more useful to improve the visibility and identification of the tide-gauge locations, for example by adding station IDs or labels. This would also facilitate interpretation of later references to specific stations; for instance, Oostende harbour is mentioned at L87 but cannot be readily identified in Fig. 1.

Section 2.3 – Methodology: In general, additional information regarding hyperparameter selection would improve the reproducibility of the study. Several architectural choices appear to be largely empirical, including the use of six Transformer layers, three LSTM layers, the selected hidden dimensions, and the number of training epochs. Please clarify whether a systematic hyperparameter search or sensitivity analysis was conducted and, if so, describe the procedure and selection criterion. The choice of a 12-hour input history also requires further justification. It would be useful to know whether alternative temporal windows were evaluated.

L117: "*we augment the gauge observations with COHERENS*."

This appears to be a central methodological aspect of the study and possibly one of its main contributions. I therefore suggest providing a more explicit description of what is meant by "augmentation" and how the COHERENS information is incorporated into the ML pipeline. At this point in the manuscript, it was not sufficiently clear to me how the observational and numerical-model information are combined.

L131–132: "*This enables the Transformer to handle long time series more naturally than long short-term memory (LSTM) networks, which must process sequences in order, and allows for efficient parallelisation during training*."

This is an interesting methodological statement. Please provide an appropriate reference supporting this comparison between Transformer and recurrent architectures.

L137–138: "*F = 583 input features: 459 atmospheric-forcing values, 102 COHERENS surge values, and 22 gauge observations*."

The model uses a relatively large number of input features (583), yet no dimensionality-reduction or feature-redundancy analysis appears to be performed. Please provide some justification for retaining the complete feature set and, if possible, discuss whether redundancy or collinearity among the predictors could influence model performance.

L173–175: "*This dataset was then split 80/20 into training and validation sets. The full year 2019, which is independent of the training period, was held out as the test set to assess the model's robustness on unseen data*."

Although the test set is independent of the training data, training the model using years chronologically later than the test year is somewhat unusual for a forecasting application. I recommend discussing the rationale for this choice explicitly. In particular, readers may question whether the experimental design adequately represents the intended operational use case, even if no direct temporal leakage occurs.

L205: "*All metrics are computed on the absolute sea level (surge residual plus astronomical tide)*."

Could the authors please justify this choice? Since the ML models are designed to predict surge residuals, it would be useful to explain why performance is ultimately assessed using absolute sea level and whether the conclusions differ when metrics are calculated directly for the surge residual.

L220–221: "*We define an extreme event as a period during which the surge residual exceeds +1.0 m or falls below −1.0 m*."

Please provide further justification for the ±1.0 m threshold. For example, is it based on a percentile of the local surge distribution, previous literature, or an operational threshold used by relevant authorities in the study area? Establishing the physical or operational basis of this definition would strengthen the extreme-event analysis.

Section 3.5 – Influence of model setting: In my view, this analysis is more closely related to model calibration or methodological sensitivity than to the main results. I suggest considering whether it would be more appropriately placed in the Methodology section or presented before the principal performance results, since the selected model configuration presumably informs the subsequent analyses.

Section 4 – Discussion: I consider the Discussion to be the section that would benefit most from further development. At present, substantial parts reiterate findings already presented in the Results. I encourage the authors to expand the scientific interpretation of the results. In particular, the manuscript would benefit from a deeper discussion of:

- The physical interpretation of why incorporating COHERENS information improves the ML predictions.
- The operational implications of the reported improvements.
- Limitations of the proposed framework and data configuration.
- Geographical and temporal transferability.
- Sources of uncertainty and their propagation through the hybrid system.
- Directions for future research.

Such discussion would help move the manuscript beyond a performance comparison and provide greater insight into when and why the proposed hybrid framework is expected to be beneficial.

TECHNICAL CORRECTIONS

L01: "*so accurate short-term predictions*". "so" sounds somewhat informal in this context. Please consider alternative wording.

L114–115: "*For example, Naeini and Snaiki (Naeini and Snaiki, 2024)*". This should be revised to "For example, Naeini and Snaiki (2024)."

Fig. 5: "COHERNS" should be corrected to "COHERENS."

L243: "*a understating*" should be corrected to "an understating."

L380: "*infomation*" should be corrected to "information."