

The paper proposes a hybrid ML framework for 2–12 h storm-surge nowcasting at 22 tide gauges in the southern North Sea. It combines gauge observations and ERA5 wind/pressure fields with gridded surge output from the operational COHERENS hydrodynamic model, feeding all three sources into a Transformer and a parallel LSTM+Transformer architecture. Trained on 2020–2024 and tested on the independent year 2019, the hybrid LSTM+Transformer achieves RMSE 0.055 m at a 2 h forecast horizon (vs. 0.164 m for COHERENS and 0.061 m for the gauge-only baseline), degrading steadily to 0.146 m at a 12 h horizon via recursive application. Sensitivity tests on look-back window length and Transformer depth are included, and the framework is proposed as an operational complement to COHERENS, refreshing the nowcast between its 12-hour update cycles.

Overall, this is an interesting study on a relevant operational problem, and the paper is well-written. The methodology is described in enough detail to follow, and comparing against both a gauge-only baseline and the operational COHERENS system is a sensible design choice. I have a number of comments below, ranging from points that need clarification to some methodological questions worth addressing before publication.

1. Absolute sea level vs. surge residual in error metrics

Table 1 and the abstract report both RMSE and R^2 on absolute sea level (surge residual plus astronomical tide) rather than on the residual itself. For RMSE this is fine: Since the ML models predict the residual and the same UTide-derived astronomical tide is added back to both observation and prediction, the tide term cancels itself out and the reported RMSE is identical to the residual RMSE. For R^2 , however, this is different. R^2 is normalized by the variance of the target itself, and on absolute sea level that variance is dominated by the tidal oscillation (order of 1–2 m amplitude), which is one to two orders of magnitude larger than the variance of the surge residual (typically a few tens of centimeters). This pushes R^2 toward 1 regardless of how well the model actually captures the surge itself. In effect, I suspect that the high R^2 values in Table 1 (0.988–0.996) and elsewhere mostly reflect that UTide reconstructs the astronomical tide very well (which is shared by all ML variants), rather than differences in the models' skill at the harder part of the problem, which is the surge residual. I would recommend reporting R^2 computed on the surge residual as the primary skill metric, alongside the RMSE already given on absolute sea level.

2. Use of meteorological forecast information

As far as I can tell from Fig. 4 and the description of the 583 input features (Section 2.3.3), every input to the model (atmospheric forcing, COHERENS surge, and gauge observations) is drawn only from the 12-hour look-back window. The input does not include the wind or pressure forecast for the period for which the surge is actually being forecasted. That is a legitimate design choice, but it means that the ML models are implicitly relying on persistence of the driving meteorology, while COHERENS itself is driven by an actual future meteo forecast, right? I think this is worth discussing directly, since it would also help explain the error growth with forecast horizon in Section 3.4: Without future forcing, degradation beyond the surge's autocorrelation timescale is expected, so it is not just an architectural weakness.

3. Generalization to extreme events

It is mentioned in the limitations, which I appreciated, but for a paper motivated by flood-risk decision support it's in my opinion worth adding some more efforts to the validation for genuinely extreme events. The two illustrated surge events (1.3 m and 1.7 m) are both within the 5-year training/test span's normal variability. Since ML models are known to extrapolate poorly beyond their training range, and at the same time, the events that matter most operationally are the ones with 10+ year return periods, it would add real value to test the trained models against more

extreme historical storms, purely as an out-of-sample test, even without retraining. Since most stations in the southern North Sea have a long observational record, this should not ask for too much extra effort, but would definitely add significant value to the verification of the method presented in this publication.

4. Number of input variables in hybrid vs non-hybrid models

The input size for the non-hybrid models is not reported, but I expect that the size is 481 (583 minus 102 COHERENS surge values). So, every hybrid variant has a larger input size than its non-hybrid counterpart, which means the hybrid models also have more trainable parameters in their first layer, independent of whether the COHERENS input itself carries useful physical information. Since additional capacity can improve fit on its own, it is currently not possible to tell how much of the hybrid models' accuracy gain (e.g. hybrid Transformer 0.096 m vs. plain Transformer 0.123 m) is attributable to useful hydrodynamic information from COHERENS, versus simply having a larger, more expressive network. I would suggest (i) reporting input parameter size for all four model configurations, and (ii) including an extra comparison, for instance a gauge-only model with its input layer artificially widened (via padding or duplicated/noised channels) to match the hybrid model's parameter count. In that way, you can isolate how much of the improvement is due to the physical content of the COHERENS input specifically, rather than model scale.

5. In 2.2.1, I miss some information on the time periods and names of the tide gauge locations. Maybe add numbers to the green dots in Fig. 1 and then a table with the names and used time periods of each station (can be in the appendix). Also I wonder why only 4 years of data (2020-2024) are used for training and validation. In my opinion this is a limitation of the study, because it undermines the representation of natural variability. Most tide gauge stations in the southern North Sea cover a much longer period, so why not use it?

6. Some of the reported differences in model performance between configurations are relatively small. With no repeated runs across random seeds, it's hard to know whether these orderings are robust or within noise. At least a few repeats (or a note on typical run-to-run variability) would strengthen the robustness of the results.

7. The extreme-event threshold (± 1.0 m surge residual) is not related to a return period or exceedance frequency, which would help readers judge how "extreme" these 185 events really are.

8. No training cost (GPU-hours) is reported, only inference time. It would be interesting for the reader to know and given that the limitations section recommends periodic retraining after major surge events, the retraining cost is operationally relevant.

9. Explain the split into training and validation with a bit more detail (is it just a random split?)

10. It would be nice to add the model domain of the ERA5 input and COHERENS model to the map in Fig 1.

11. Section 4.1: No experiment isolates a COHERENS-only input (i.e. without gauge observations). This would help clarify how much of the models' skill is transferable to ungauged or short-record stations. This could be a case for which this framework could be well-suited, since it wouldn't require a long local observational record.

12. Explain in more detail what is meant by "Positional encoding" in Fig. 3.

13. Have you performed any other sensitivity analyses on the model architecture, like dropout rate, amount of training epochs, learning rate etc.? Would be informative to explain your model architecture choices in Section 2.3.4.

Minor corrections

- Section 3.5: "sensitivity of the hybrid LSTM+Transformer **agaisnt** two key settings" → "against".
- Section 4.1: "the hybrid **LST+Transformer** still yields the best performance" → "LSTM+Transformer".
- Section 3.3: "Apart from the surge residual, it is also important to have **a understating** of" → **an understanding** of (?)
- Model naming is inconsistent in Section 3.3/3.4 ("LSTM–Transformer" vs. "LSTM+Transformer" elsewhere).
- Some of the Figure and Table captions could be more descriptive/standalone (e.g. tables 2 and 3 are for a forecast horizon of 2 hours ahead?).
- Figure 8 (panel c): The color choice for the two look-back-window variants (6 h vs. 12 h) is hard to distinguish, the magenta/red pairing is difficult to tell apart at a glance. Consider using more visually distinct colors to make the two series easier to separate.