Response to reviewer

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.

Answer:

We sincerely thank reviewer for the thorough and constructive assessment of our manuscript. These valuable comments and suggestions have helped us to significantly improve the quality of the manuscript. Our responses to the individual comments detail the scientific and methodological enhancements that will be implemented in the revised manuscript.

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.

Answer:

Thanks for your detailed comments and we agree with you for the matter of error metrics computation. We will provide the R^2 values for surge residuals alongside the R^2 values for absolute sea level in the new table in our revision.

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.

Answer:

Thanks for this very interesting comment.

Meteorological forecast are the main forcing of the COHERENS model. This means that (i) at each time step, the state of the COHERENS model results from the influence of the past weather conditions, and (ii) when updating the sea conditions at the next time step, the COHERENS model is only informed of the current weather conditions. In that sense, COHERENS is not informed of the weather conditions at time steps that have not been computed yet.

This being clarified, we agree with the suggestion of the reviewer : the input vector of our ML models might be extended to incorporate the future meteorological and surge states as available from ERA5 and COHERENS. This might be the topic of a future study.

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.

Answer:

We fully agree that machine learning models generally exhibit poor extrapolation capabilities beyond the range of conditions represented in their training data. Ideally, the training period should therefore encompass the longest possible time span, particularly considering that the most severe storm-surge events may have return periods of 100 years or more. For example, the Belgian coastal sea-defense system has been designed to withstand storm surges with a return period of 1,000 years (MDK and WL, 2011, p. 22).

In this context, the five-year training period used in the present study may appear relatively short. However, the objective of this article is not to develop an operational forecasting system, but rather to demonstrate the feasibility of hybrid ML approaches for storm-surge nowcasting applications. In other words, the study aims to investigate and assess different model architectures and evaluate their potential for this type of application. Importantly, the results presented in Section 3 demonstrate that, even when trained on only five years of data, the proposed models generalize well to typical storm events.

The most obvious next step would be to train the models on the longest possible historical records. However, contrary to the reviewer's suggestion, this would require substantial additional effort. First, a consistent multi-decadal storm-surge hindcast would need to be produced using the COHERENS model, which would constitute a dedicated research project itself. Second, and perhaps more importantly, the availability and quality of historical observations vary considerably among tide gauges. This heterogeneity represents a major limitation for the multi-station framework adopted in the present study.

Agentschap Maritieme Dienstverlening en Kust (MDK) and Waterbouwkundig Laboratorium. (2011). Masterplan Kustveiligheid. Vlaamse Overheid. 90pp. https://www.agentschapmdk.be/nl/bijlage/0cd6feb8-e135-4c01-8057-cb206772ea20/masterplan-kustveiligheid.pdf

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.

Answer:

We totally agree with your suggestion, and it is valuable for further explaining our study. The number of parameters for Transformer, hybrid Transformer, LSTM+Transformer and hybrid LSTM+Transformer are 0.37 M, 0.38 M, 21.63 M, 23.75 M (corresponding number of parameters are 97364, 100628, 5,668,746, 6223988) for float 32 respectively. We will include this as well as the four models' training time in our table 3 when we report the error metrics for the four models. Meanwhile, we will add another experiment "*add an extra comparison by padding the noised channels for the LSTM + Transformer model with tide gauge observation only into same input size of hybrid LSTM+Transformer model*" and report its performance as the reviewer suggested.

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?

Answer:

Fig 1. Shall be updated according to your suggestions. In particular, the stations will be clearly labelled with counter-clock order in the following example. We agree with the reviewer and in table 1 of appendix A, we will show the station IDs and names in column 1 and 2 to enable readers to find the stations correspondingly.

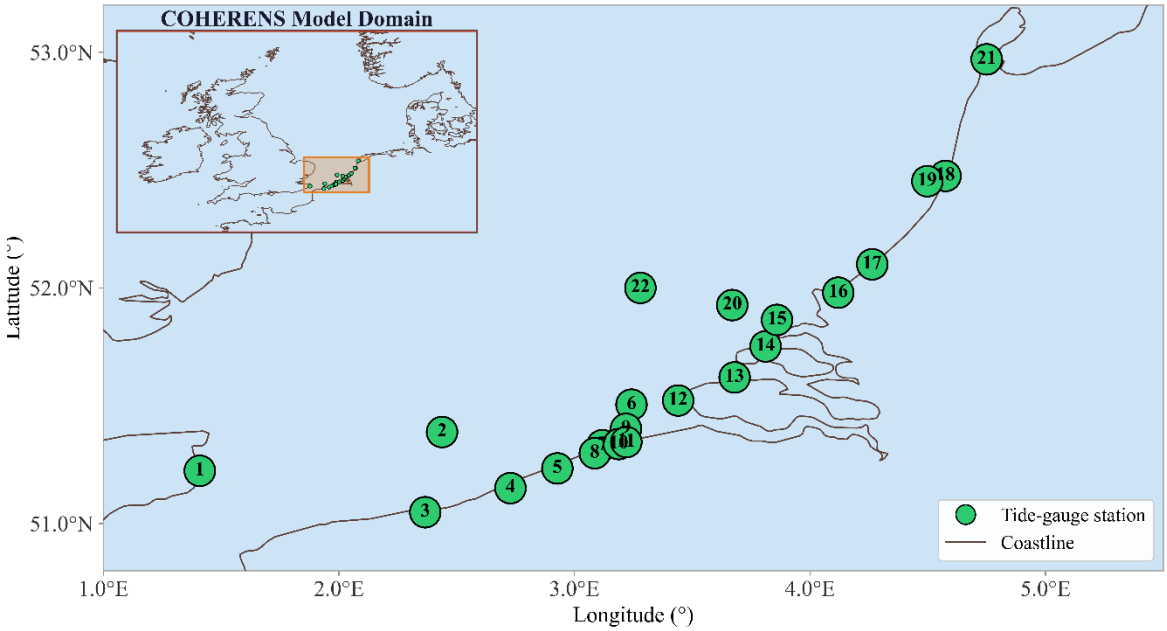


Figure 1. Distribution of the tide gauge stations (green dots) over the southern bight of the North Sea

The 5 years training periods (not 4) have been selected mainly as a constrained on the availability of the COHERENS data set.

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.

Answer:

We agree with the reviewer's comments and will repeat the experiment for 5 different random seeds. The revised manuscript, we will report and discuss the variances generated by the different model configurations.

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.

Answer:

The ±1 m threshold was originally defined in consultation with an expert in operational oceanography. We acknowledge that this choice is somewhat arbitrary and, in the revised manuscript, we may replace it with a data-driven criterion. For example, the 99th percentile of surge values over the 2020-2024 reference period corresponds to thresholds of -0.52 m and +0.75 m for negative and positive surges, respectively.

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.

Answer:

In the revised manuscript, Table 3 will be extended to give the training time along with the error metrics of the different model configurations.

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

Answer:

Section 2.4 will be revised to clearly explain that the dataset was divided into training and validation sets following an 80/20 split based on chronological order, thereby preventing information leakage from future observations into the training phase.

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

Answer:

Figure 1 will be revised in accordance with the other reviewer's comments as well as your previous and this suggestion.

.

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.

Answer:

We had trained a hybrid LSTM+Transfomer model with the COHERENS-only input, but as expected, the validation metrics of this model remains similar to the validation metrics of COHERENS. This is the reason why the test has not been reported in the article. However, we can report its performance in the revised manuscript in table 10 (row COHERENS solely).

Table 10.Error metrics of COHERENS solely and extended gauge observation (replicate channel with noise) with forecast horizon of 2 hours

Model	RMSE (m)	R^2 for surge residual (-)	R^2 for absolute sea level (-)	Bias (m)
COHERENS solely	0.143	0.596	0.970	-0.024
Extended gauge (replicate channel with noise)	0.054	0.939	0.996	-0.006

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

Answer:

Positional encoding is a technique used in transformer neural networks to provide information about the order of tokens in a sequence. This technique preserves information about the ordering of the input vectors while they are processed by the transformer. In this work, the standard fixed sinusoidal positional encoding has been adopted. To make this choice transparent to the reader, Figure 3 will be updated in the revised manuscript.

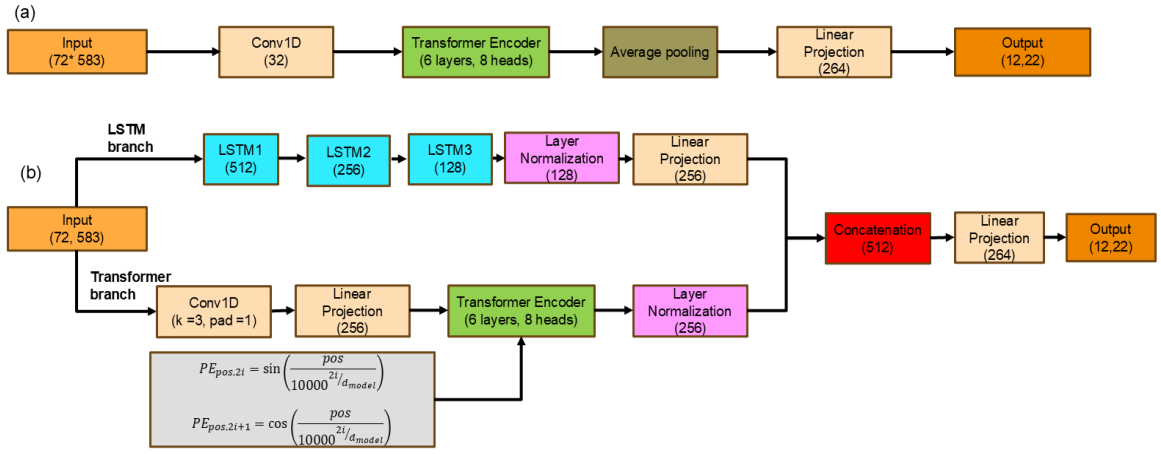


Figure 3.Architecture of (a) the hybrid Transformer and (b) the hybrid LSTM+Transformer model

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.

Answer:

In this manuscript, we conducted targeted ablation studies to validate key architectural choices. Specifically, we varied the number of Transformer layers from 5-8 and LSTM layers from 2–4 on our LSTM+Transformer model configuration. Regarding the hidden dimensions, we also tried a few, including dimesons of [128,64,32], [256,128,64], [512, 128, 64] and [1024, 512, 256]. Apart from it, we also tried to vary the learning rate with the number of 0.00001,0.00002,0.00005, 0.002,0.005 as well as the dropout rate with 0.05 and 0.2. The error metrics of corresponding experiments based on the validation dataset in the training procedure are summarized in the following table.

Table 2. Error metrics of hybrid LSTM +Transformer by using various hyperparameters

		RMSE (m)	R^2 (surge residual)	R^2 (total height)	Bias (m)
Baseline (6 Transformer layers, 3 LSTM layers, Learning rate of 0.0001, hidden dimensions [512, 256, 128])		**0.059**	**0.935**	**0.996**	**0.012**
Transformer layers	5	0.065	0.928	0.994	-0.013
	7	0.058	0.937	0.996	0.022
	8	0.066	0.918	0.995	0.026
LSTM layers	2	0.066	0.924	0.993	-0.003
	4	0.062	0.933	0.995	-0.019
drop out rate	0.05	0.057	0.942	0.996	-0.004
	0.2	0.059	0.938	0.995	-0.003
Learning rate	0.00002	0.057	0.941	0.996	0.015
	0.00005	0.054	0.946	0.996	-0.002
	0.0002	0.060	0.935	0.995	0.004
	0.0005	0.083	0.885	0.990	-0.005
Hidden dimension	[256,128,64]	0.076	0.902	0.992	-0.012
	[512, 128, 64]	0.057	0.938	0.996	0.010
	[1024, 512, 256]	0.072	0.913	0.993	-0.003

It can be shown that by varying the hyperparameters slightly, the validation performance doesn't change dramatically (with the RMSE with respect to the baseline hyperparameter setting is less than 0.01 m). This implies that the chosen values are in a stable region, which is reasonable for later application. The training epochs were determined by early stopping with a patience of 11 based on validation loss, making that choice principled rather than purely empirical. After checking the training records, the training stops at 50- 80 epochs, so 200 epochs we chosen is sufficient. We acknowledge that a full hyperparameter sweep was not performed due to computational constraints. The revised manuscript will incorporate this clarification.

Minor corrections

Section 3.5: "sensitivity of the hybrid LSTM+Transformer agaisnt two key settings" → "against".

Answer:
The typo shall be corrected in the revised manuscript.

Section 4.1: "the hybrid LST+Transformer still yields the best performance" → "LSTM+Transformer".

Answer:
The typo will be corrected in the revised manuscript.

Section 3.3: “Apart from the surge residual, it is also important to have a understating of” → an understanding of (?)

Answer:
The typo shall be corrected in the revised manuscript.

Model naming is inconsistent in Section 3.3/3.4 (“LSTM–Transformer” vs. "LSTM+Transformer" elsewhere).

Answer:
We will double check the model naming in the revised manuscript to make sure it is consistent throughout the manuscript.

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?).

Answer:
Figure captions will be made more descriptive in the revised manuscript.

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.

Answer:
A green/red color pairing will be used in the revised manuscript.