Reviewer 1

The manuscript presents an approach to storm surge prediction using deep learning. Several methodological issues must be addressed to improve the manuscript's clarity and strength. Specifically, clarifying dataset construction, justifying hyperparameter choices, and improving performance evaluation will significantly enhance the manuscript. In its current form, the manuscript is not appropriate for publication.

We thank the reviewer for scrutinizing our methods. As further detailed below, we will address the reviewer's comments by better motivating the methodological choices that we made, providing additional details and discussion of aspects that were unclear before, and performing several tests to better justify the model parameters that we used. While we agree these revisions will strengthen our manuscript, our results and conclusions still apply.

Introduction:

L41-48: The last phrase "Furthermore, because several … hydrodynamics models" does not sound logical to me. It sounds like the authors reached a "general" conclusion from the previous "several" studies.

We will replace *"several"* by *"most"*, and restructure the sentence to "*Furthermore, how neural networks compare to state-of-the-art hydrodynamic models in this regard also remains unclear, because most previous studies either did not specifically evaluate the extremes or considered extremes exceeding relatively low thresholds (e.g., Bruneau et al., 2020; Tadesse et al., 2020; Tiggeloven et al., 2021)*" to improve its clarity.

Methodology – Data Preparation:

Dataset Size and Class Distribution: The paper mentions using data from 1979 to 2017 at a three-hour resolution. However, it is unclear how many training samples remain after filtering or how the extreme events (99% and 99.9%) are distributed.

Thank you for pointing this out. We will add a new table containing the total number of samples, the magnitude of the 99th and 99.9th percentiles, and the number of filtered extremes exceeding these percentiles for each tide gauge and split (Table B1, copied below). References to the new table will be added in Sections 2.4, 3.3 and 3.4. We will also clarify why the split sizes presented in Table B1 deviate slightly from the nominally defined split-size ratios in Section 2.3 by adding: *"Due to differences in tide-gauge data coverage between splits, the true split-size ratios can deviate from the nominal ones by up to a few percent (see Table B1)."* to L96.

Table B1. Number of samples, the magnitude of the 99th and 99.9th percentiles (P_{99} and $P_{99.9}$) [m], and the number of filtered (see Section 2.4) extremes exceeding P_{99} and $P_{99.9}$, per split and per tide gauge.

	Samples [#]			P_{99}^{split} [m]			$\geq P_{99}^{split}$ [#]			$P_{99.9}^{split}$ [m]			$\geq P_{99.9}^{split}$ [#]		
Tide gauge	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test	Train	Val	Test
1 Stavanger	65094	23178	21697	0.35	0.35	0.35	586	199	190	0.53	0.51	0.55	66	24	22
2 Wick	57832	21538	16200	0.43	0.43	0.41	516	192	138	0.63	0.63	0.61	59	22	16
3 Esbjerg	62889	20249	21364	1.07	1.09	1.10	572	187	192	1.85	1.69	1.72	63	21	22
4 Immingham	52939	20280	20164	0.56	0.57	0.56	403	153	159	1.01	0.93	1.00	51	20	20
5 Den Helder	66485	23368	23376	0.81	0.80	0.80	593	206	200	1.40	1.26	1.39	67	24	24
6 Fishguard	57995	18361	19284	0.38	0.39	0.39	475	153	156	0.59	0.64	0.65	55	19	18
7 Brest	67810	21748	23304	0.35	0.35	0.36	577	168	195	0.55	0.53	0.59	65	19	23
8 Vigo	59772	22775	22663	0.29	0.29	0.30	487	168	207	0.45	0.42	0.45	53	16	23
9 Alicante	53578	16692	19681	0.20	0.20	0.20	493	154	182	0.29	0.28	0.31	56	16	19

Explanatory Variables: While the paper includes zonal, meridional, and absolute wind speed as predictors, absolute wind speed is directly derivable from the other two. The authors need to justify this inclusion. Otherwise, removing absolute wind speed could prevent redundancy and improve efficiency.

We agree that absolute wind speed is a deterministic transformation of its zonal and meridional components. We initially included it based on prior work (Tiggeloven et al., 2021), who found that adding physically meaningful, deterministically derived predictor variables can sometimes improve model performance. Extensively testing the sensitivity of our models to the predictor variables was not our focus, but to still evaluate this we carried out an ablation study at one of the tide gauges (Esbjerg). This involved training the LSTM model with three different predictor sets: (1) only sea-level pressure, (2) sea-level pressure and zonal and meridional wind, and (3) sea-level pressure, zonal and meridional wind, and absolute wind speed (the default). We will include these experiments in a new appendix (Appendix A) and have copied the results below for the reviewer's convenience.

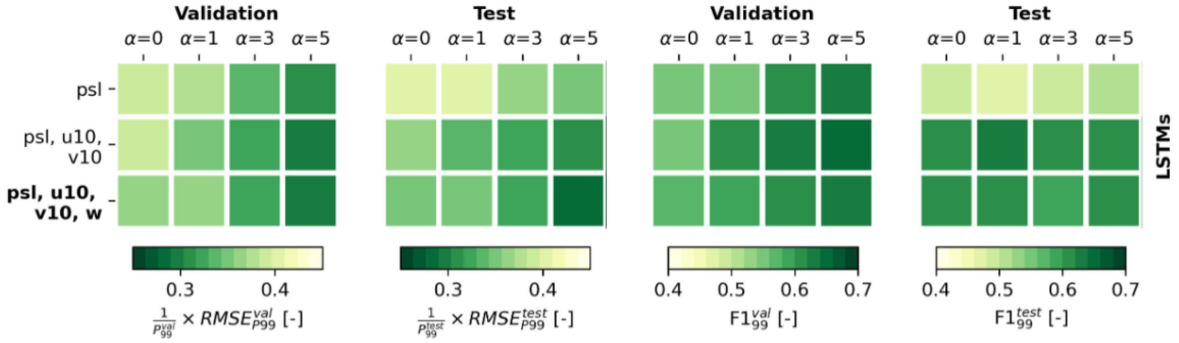


Fig. A1 (top row). Sensitivity of the average $RMSE_{P99}$ relative to P_{99} [-] and $F1_{P99}$ [-] of the LSTM and ConvLSTM models at Esbjerg (DK) to the predictor variables (mean sea-level pressure psl, zonal and meridional wind u10 & v10, and absolute wind speed w10), for different values of α. The error metrics are shown for both the validation (1st and 3rd columns) and the test splits (2nd and 4th columns). The bold text on the left of the figure indicates the default settings used for the results in the main manuscript.

The results show that using the zonal and meridional wind components in addition to sea-level pressure clearly improves model performance, but that additionally using absolute wind speed does not. Importantly, however, using absolute wind speed does not clearly degrade overall model performance and generalization either. Based on these results, we conclude that absolute wind speed may indeed not be necessary, although we cannot exclude potential benefits at other locations. Since it did not degrade model performance at Esbjerg either, we therefore opt to keep absolute wind speed as a predictor variable at all sites. The detailed discussion of Figure A1 above will be added to the new Appendix A, and to clarify this in the manuscript, we will add the following:

"*Absolute wind speed was included based on previous research that showed that including derived but physically meaningful predictor variables can provide added value (Tiggeloven et al., 2021). Our sensitivity tests at the tide gauge in Esbjerg, however, suggest only a minor influence (see Appendix A). To improve model efficiency, future work could therefore investigate whether absolute wind speed could also be left out without substantially impacting model performance at other locations.*" to L83.

Construction of Data Points: Storm surges can persist for several days. It is essential to clarify whether data points overlap, whether each event is treated as an independent sample, or if multi-day storm surges are captured uniquely.
Threshold exceedances are treated independently to maximize the number of extreme samples available and allow the model to learn about the temporal evolution of storm surges. In other words, exceedances can be part of the same storm, and the predictor data used for them can partially overlap because of the 24h look-back window. As we agree these are important points to clarify, we will:

1) add *"Because of the look-back window, the predictor data used for predictions at consecutive 3-hourly time steps partially overlap."* at the end of L85.
2) add *"We treated remaining threshold exceedances independently regardless of whether they occurred during the same event, because this allows the neural networks to learn about the temporal evolution of storm surges, and uniquely capturing storm surges through declustering would reduce the available sample size unless more moderate events would be considered.The numbers of filtered exceedances in each split are shown in Table B1."* to L147.

Atmospheric Variables: The authors predict sea-level height based on ERA5 atmospheric data but do not specify whether land-based data is included. If land data is incorporated, the authors need to justify and discuss how it was handled.
As shown in Figure 1, the atmospheric predictor data includes values over land. We did not exclude this data because, as part of an atmospheric pattern over a given location that typically extends hundreds of kilometers, it may provide useful additional information for the prediction of storm surges. To clarify, we will add "*The predictor data includes grid cells over land, which do not directly affect water levels, but, as part of a certain weather pattern over a location, may contain features relevant for predicting storm surges.*" to L84.

Model Training & Hyperparameter Tuning:

Training Epochs: The authors use a maximum of 100 training epochs with early stopping. Given the complexity of the models, 100 epochs may not be sufficient. The authors need to justify the convergence of the model with 100 epochs.
We thank the reviewer for this comment as it prompted us to better motivate this choice. It was not feasible to visually inspect the loss evolution of all experiments, but we find that the maximum number of training epochs was reached in only 6.2% of all LSTM experiments and 5.6% of all ConvLSTM experiments, primarily when the lowest learning rate (1e-5) was used. In those instances, the average decrease in the validation loss over the last 5 decrements during training was only two to three tenths of a percent, with each decrement typically requiring multiple epochs. We therefore do not expect more training epochs to substantially improve our results.

We will add this justification to Section 2.3: *"The maximum number of epochs was reached for only 5 to 6% of all LSTM and ConvLSTM models, and based on the small decrements in the validation loss near the end of the training of these models, we do not expect that using a maximum of more than 100 training epochs would substantially improve our results."* (L130).

Dropout Rate: The dropout rate of 0.1–0.2 may be too low. LSTM and ConvLSTM models often use dropout rates of 0.3–0.5 to prevent overfitting. If different dropout rates have been tested, discussing their impact would improve transparency.
Upon comparison, we find that the agreement between the average performance metrics in the validation and test splits does not structurally improve with a dropout rate of 0.2 compared to a dropout rate of 0.1, suggesting that further increasing the dropout rate is not necessary to prevent overfitting and improve generalization in this case. To clarify, we will add the following note to Section 2.3 (L135): "*As a dropout rate of 0.2 did not lead to a structurally better generalization of the models to the independent test split than a dropout rate of 0.1, we did not increase the dropout rate beyond 0.2.*".

Performance Evaluation:

Evaluation Metrics: The authors use the F1-score as a primary evaluation metric. In extreme event prediction, recall is often more important than precision, as missing a storm surge event is more consequential than a false positive. A high F1-score does not necessarily indicate strong model performance if recall is low. Reporting recall and precision alongside the F1-score would provide a more comprehensive assessment. A confusion matrix could be also beneficial.
We agree that this may be the case in the context of flood forecasting, but for fitting extreme value statistics and projecting long-term changes in extremes, false positive predictions due to a lower precision would also adversely affect results, which is why we chose to use the F1-score alongside the RMSE. To better reflect this, we will change L199-200 *"Namely, increasing the density-based weights generally leads to more true positives but also to more false positives"* to "*Namely, increasing the density-based weights generally leads to a higher recall but a lower precision (i.e., less false negatives but also more false positives)*" and explicitly mention that the decreasing precision of neural networks trained with a higher α parameter may "*negatively influence subsequent extreme-value analyses.*" (L194). We

agree these aspects are important, which is why the recall and precision as a function of α are shown for the different locations in Fig. C1 (will become Fig. D2), which is referred to in Section 3.1. To acknowledge that in some contexts, recall may be more important than precision, we will add to the discussion in Section 4 that the hyperparameter α needs to be tuned *"in relation to the problem context. For instance, a higher α value may be better for applications in which recall is more important than precision (see Fig. D2)."* (L281).

Discussion & Conclusion:

The discussion and conclusion are well-written based on the current results of the study, but they will need to be updated after the revision of the manuscript.
We thank the reviewer for their positive feedback on our discussion and conclusions. Our revisions in response to the reviewer's comments will mainly involve improving the explanation and justification of our methods. As detailed above, major updates to the discussion and conclusion sections were not necessary.

Minor Edits:

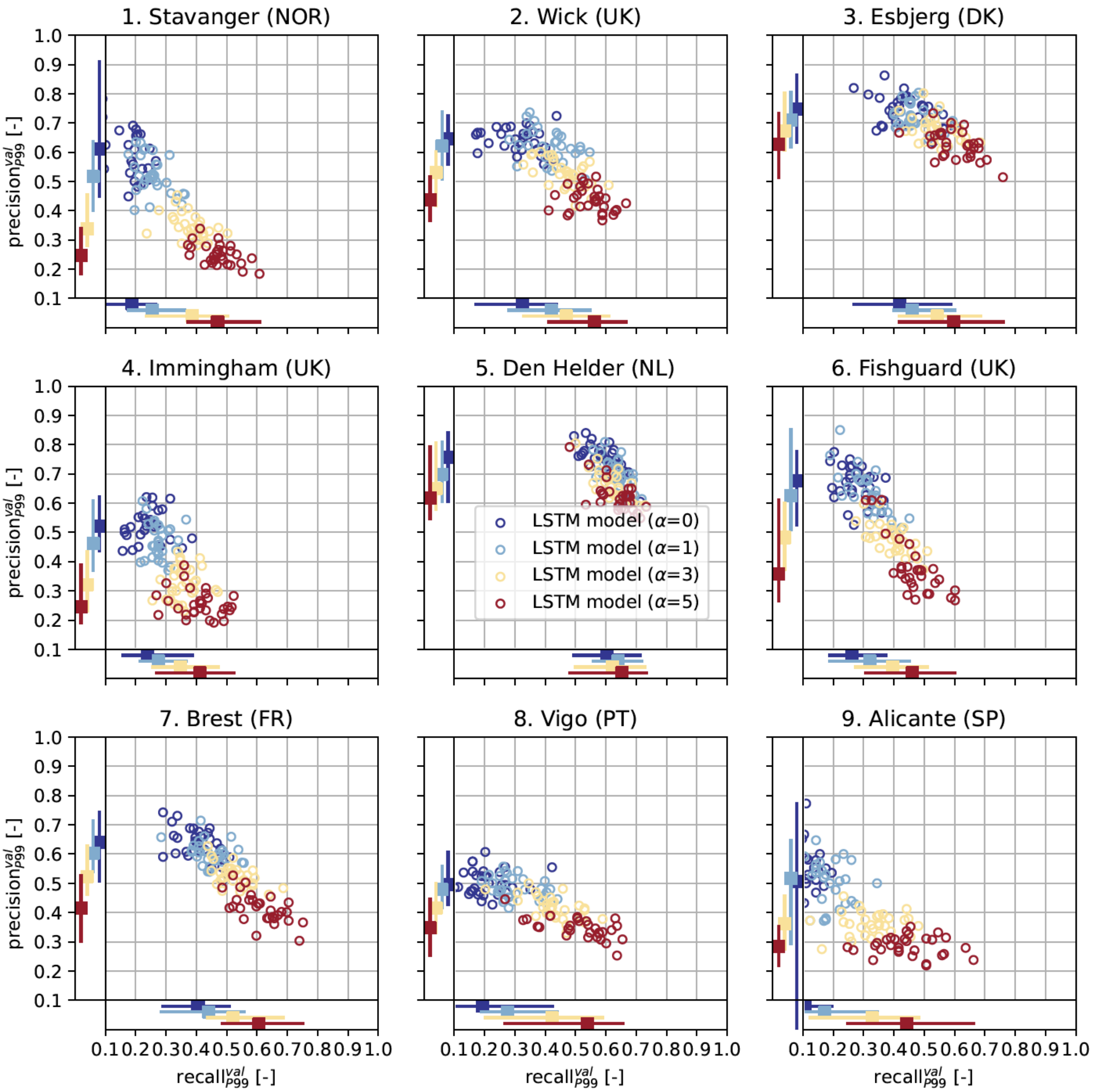
"v.s." to "vs."
This will be corrected throughout the paper.

L41: "more moderate" to "moderate."
Thank you. We will remove *"more"*.

L318: Remove "at least."
We will replace "*at least*" with *"especially"*.

Additional changes (to be appended to both responses)

1. We discovered a minor error in the stratification routine introduced in Section 2.1, where 1 year of input data was accidentally omitted at the locations Wick and Alicante. We fixed this error, trained the models again at these locations and will update the corresponding results throughout the paper where needed. Fixing the error led to only minor differences in the results.
2. Figure C1 (now D2) accidentally showed results for the test split, instead of for the validation split. We apologize for this mistake, and will replace the figure with its correct version (see below). As the general effect of α on precision and recall is consistent across the validation and test splits, the discussion of Figure C1 (now D2) in Section 3.1 will not need to be changed.



Corrected version of Fig. C1

3. Figure 4 aggregates results for all locations, which we realized makes it harder to inspect model generalization at individual locations. We will therefore complement Figure 4 with an additional supplementary figure (Figure D3) that shows the same results, but separated by location (see below).

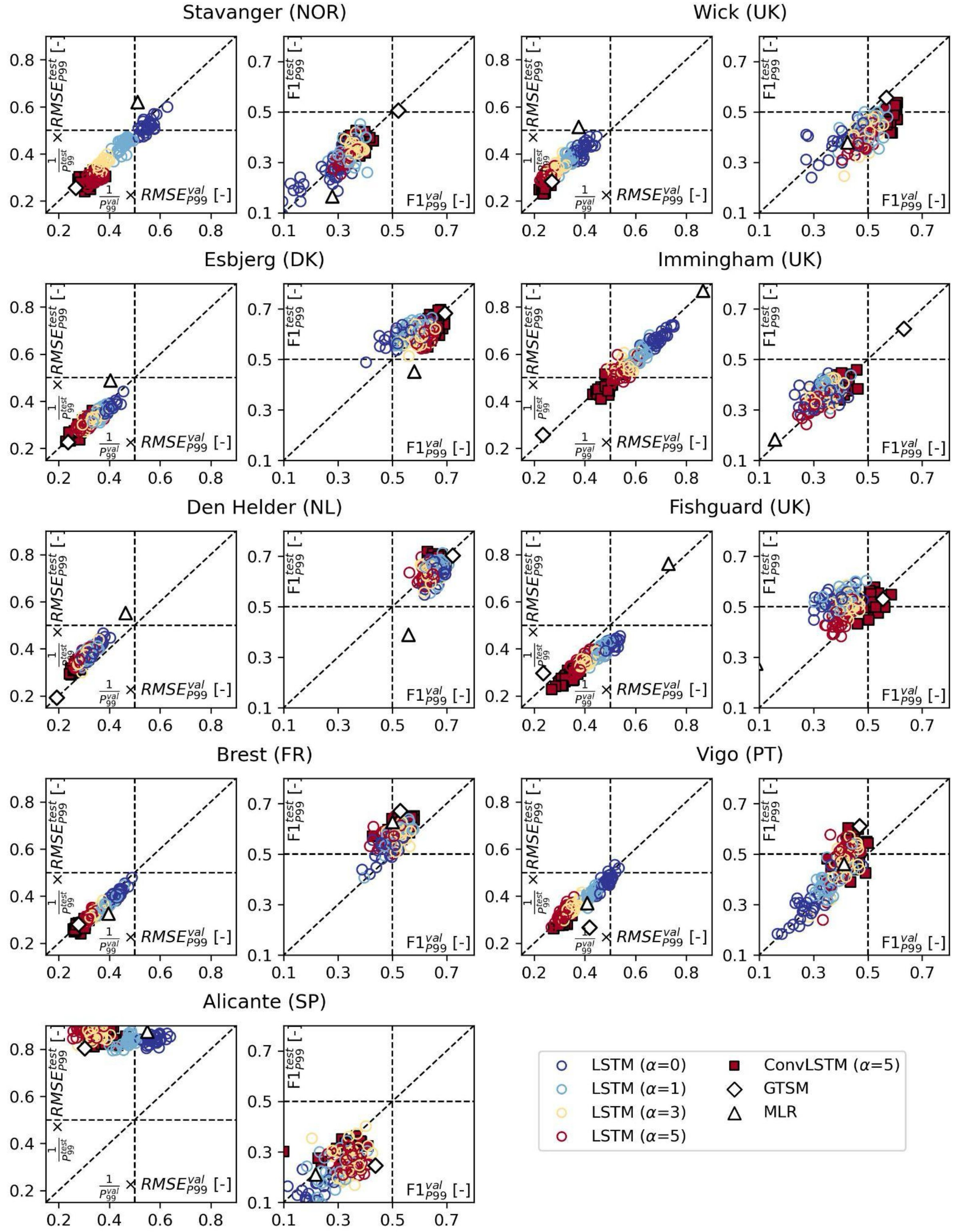


Fig. D 3. Scatter plots of $RMSE_{P99}$ relative to P_{99} [-] and $F1_{P99}$ [-] in the validation vs. in the test split, displayed per tide gauge. The colored circles represent the LSTM models for different values of α, the black-edged red squares the ConvLSTM models for α=5, and the white triangles and diamonds the MLR model of Tadesse et al. (2020), and GTSM (Muis et al., 2020), respectively. The diagonal line in each panel indicates equal error metrics in the validation and test splits.

A few lines of discussion of Figure D3 will be added to Section 3.3: "*Additionally, we find that increasing α leads to a lower $RMSE_{P99}$ in both the validation and test splits (Figure D3)*." (L241), *"With a few exceptions, increasing α has an approximately similar effect on $F1^{test}_{P99}$ as it has on $F1^{val}_{P99}$ (Figure D3)."* (L245), and "*For optimal model generalization, we therefore recommend tuning α alongside other important (hyper)parameters using k-fold cross-validation*." (L248).