Dear editor, dear reviewers.

We appreciate your encouraging comments and thorough review of our work. Below is our point-by-point response to the second report, along with our revisions based on the reviewer's comments.

COMMENT 1:

The authors have thoroughly revised the manuscript, significantly enhancing its clarity. However, one comment related to Table 1 remains, originating from an unclear description in the original manuscript. Additionally, I have three suggestions connected to this comment.

1. For Comments 6 and 10: In my initial review, I misunderstood that HIDRA3 could handle a 72-hour time series with some missing values by filling the gaps. For this reason, I also thought the SSH availability reported in Table 1 represented the total number of scenarios used to train and evaluate the model. However, after reviewing the authors' response and the revised manuscript, it appears that HIDRA3 excludes any tide station data if there is even one missing data point in the 72-hour series. Consequently, the SSH availability in Table 1 may overstate the actual number of usable scenarios, as only a subset of cases contains a complete 72-hour time series. If this interpretation is correct, I recommend the authors clarify the exact sample size used for model training and testing to ensure readers understand the number of scenarios ultimately included.

RESPONSE 1:

We agree with the reviewer that Table 1 should reflect the sample sizes used for training and testing. We updated the table with the new SSH availability. For stations with SSH signals without interruptions, the numbers remain unchanged.

The new table (not latex-diff) and the text changes (latex-diff) are displayed below:

2.1 HIDRA3 training and testing datasets

Our objective is to forecast hourly SSH values for $N = 11$ tide gauges located along the Adriatic coast (Fig. 1) over a three-day period. HIDRA3 achieves this by leveraging a comprehensive set of ocean state parameters. This includes the past 72 hours of available sea level observations from stations shown in Fig. 1, with data availability for each station detailed in Table 1. ~~Additionally, it~~ When calculating the availability, only SSH measurements with 72 preceding measurements available are considered, as required for HIDRA3 input. Besides past SSH measurements, HIDRA3 considers both past and future astronomic tides at ~~these~~ the stations, and past and future 72 hours of gridded geophysical variables from atmospheric and ocean numerical models.

Location	SSH Availability in 2000–2022	Thresholds [cm]
Koper	90.8%	-69.3, 65.7
Venice	64.6%	-64.3, 61.3
Ancona	50.4%	-39.9, 44.6
Ortona	45.3%	-34.0, 39.6
Vieste	44.9%	-33.3, 36.5
Neretva	38.9%	-32.6, 37.8
Ravenna	37.7%	-56.3, 57.2
Sobra	24.1%	-33.4, 37.0
Stari Grad	23.9%	-34.0, 38.7
Tremiti	18.2%	-32.4, 37.0
Vela Luka	16.6%	-31.9, 38.6

Table 1. Availability of SSH measurements between 2000 and 2022 for 11 tide gauge locations used in training and evaluating HIDRA3, and defined thresholds [1st, 99th percentile] for low and high SSH values used in this study. When calculating SSH Availability, only SSH measurements with 72 preceding measurements available are considered, as required for HIDRA3 input. See Fig. 1 for station locations.

COMMENT 2:

2. Although these additional minor comments were not raised in the first review due to misunderstanding the SSH data and Table 1, I believe they would help improve the manuscript.

2.1. For Figure 3, how was this plot created if each station had a different amount of available data? Was only the overlapping data for all stations used? If so, please clarify whether this approach is valid.

RESPONSE 2:

We used overlapping data for each pair of locations individually. These differences, shown in Figure 3, estimate the increase in MAE if we were to forecast SSH at location A with some model and then apply that forecast to location B without any modifications. While the amount of data used varies for each pair, our primary goal is to accurately estimate the increase in MAE. Therefore, we decided to utilize as much data as possible, prioritizing the correctness of the scores themselves rather than the comparability of different scores. We have revised the manuscript as follows:

of each other and thus exhibit similar SSH phases of high and low sea levels. This is illustrated in Fig. 2, which depicts the SSH at all stations, and in Fig. 3, which shows mean absolute differences between all stations. These mean absolute differences were calculated using overlapping data from each pair of locations, and they can be interpreted as estimates of the increase in mean absolute error (MAE) when applying some model's forecast from one location to another.

Figure 3. Mean absolute differences [cm] of SSH measurements between different tide gauge locations. These differences estimate the increase in MAE when applying some model's forecast from one location to another. Abbreviations used here are: KP - Koper, VE - Venice, RA - Ravenna, AN - Ancona, OR - Ortona, TR - Tremiti, VI - Vieste, SO - Sobra, VL - Vela Luka, NE - Neretva, and SG - Stari Grad.

COMMENT 3:

2.2. Figures 10, 14, 17, and 19: Based on Figures 2 and 3, it appears that Koper, Venice, and Ravenna have larger MAE values than other stations due to their wider water level ranges. Additionally, each station may have a different number of test scenarios based on the SSH availability reported in Table 1. This variation raises questions about whether the MAE accurately reflects performance at each station. I suggest presenting normalized error statistics instead. Please also clarify if different sample sizes were used to calculate the error statistics and if comparing these across stations with different sample sizes is valid.

RESPONSE 3:

We agree with the reviewer that the higher MAE scores observed in northern locations are related to greater variances in SSH in those regions. To expose this, we calculated the normalized MAE (nMAE) scores and presented them in Fig. 11. This plot provides additional insights: even after normalization, NEMO shows the highest errors in northern locations. In contrast, HIDRA2 exhibits the highest normalized errors in southern locations, likely due to lower data availability in those areas. We have also calculated the mean nMAE scores and included them in Tables 2, 3, and 4.

However, it is still important to include MAE scores without normalization. Potential users of HIDRA tend to be more interested in MAE scores expressed without normalization, as these scores have immediate practical significance. Additionally, MAE without normalization more effectively reflects flood forecasting capability, since higher SSH variability indicates a greater likelihood of flooding.

Additions made to the manuscript:

3.2 SSH forecast performance

The following performance measures (Rus et al., 2023) are employed: mean absolute error (MAE), root mean squared error (RMSE), accuracy (ACC), bias, recall (Re), precision (Pr) and F1 score. Additionally, we calculate the normalized mean absolute error (nMAE) by dividing the MAE score for each location by the standard deviation of all historic SSH measurements for that location. These performance metrics are reported in Table 2 separately for all SSH values (*overall*) and for low and high SSH values (see Sect. 2 for the definitions).

	Model	MAE (cm)	nMAE	RMSE (cm)	ACC (%)	Bias (cm)	Re (%)	Pr (%)	F1 (%)
Overall	NEMO	2.65	0.142	3.56	97.76	-0.31	/	/	/
	HIDRA2	2.63	0.146	3.56	98.15	-0.17	/	/	/
	HIDRA3 (ours)	**2.42**	**0.134**	**3.28**	**98.60**	**-0.00**	/	/	/
Low SSH Values	NEMO	4.19	0.215	5.23	92.91	2.88	94.04	**99.92**	96.39
	HIDRA2	**3.27**	**0.175**	4.27	95.94	**1.02**	97.64	99.55	98.51
	HIDRA3 (ours)	3.30	0.177	**4.24**	**96.16**	1.33	**98.04**	99.85	**98.88**
High SSH Values	NEMO	4.68	0.244	6.19	89.14	-3.02	94.53	**99.40**	96.79
	HIDRA2	4.80	0.266	6.53	89.49	-2.35	96.62	97.82	97.18
	HIDRA3 (ours)	**4.06**	**0.220**	**5.61**	**91.63**	**-2.06**	**97.58**	98.67	**98.09**

Table 2. Performance calculated on all SSH values, low SSH values and high SSH values, averaged over all locations. The proposed HIDRA3 has the best performance overall and on high SSH values, and a comparable performance on low values to HIDRA2.

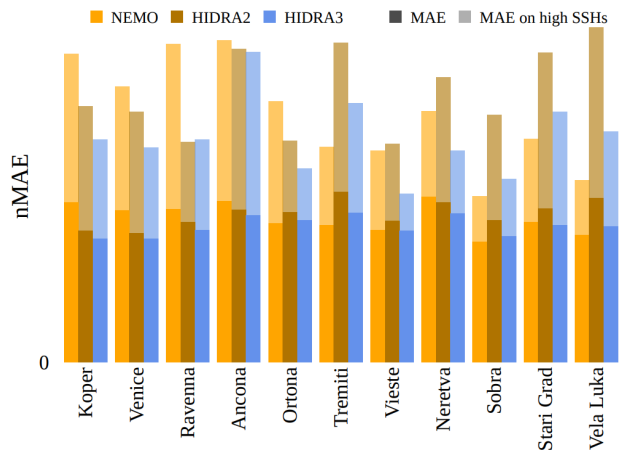

Figure 11. The normalized MAE (nMAE) calculated for all SSH values and for high SSH values, across different models and tide gauge locations. HIDRA3 demonstrated the most consistent performance, significantly outperforming NEMO (Madec, 2016) at northern locations (Koper, Venice and Ravenna), and HIDRA2 (Rus et al., 2023) at other locations.

To enable a more effective comparison of errors across different locations, we present the normalized MAE scores (nMAE) in Fig. 11. Although the scores are normalized, NEMO still shows the largest errors at northern locations (Koper, Venice and Ravenna). In contrast, HIDRA2 records larger normalized errors at the southern locations, likely due to lower data availability in those areas (see Table 1). HIDRA3 demonstrates the most consistent performance, significantly outperforming NEMO in the northern locations and HIDRA2 in the southern locations. On average, HIDRA3 has a lower nMAE score than both NEMO and HIDRA2 when calculated on all SSH values and high SSH values (see Table 2).

	Model	MAE (cm)	nMAE	RMSE (cm)	ACC (%)	Bias (cm)	Re (%)	Pr (%)	F1 (%)
Overall	$NEMO_0$	3.26	0.173	4.15	95.81	**0.03**	/	/	/
	HIDRA3 (ours)	**2.63**	**0.146**	**3.52**	**98.35**	-0.07	/	/	/
Low SSH Values	$NEMO_0$	4.00	0.217	4.95	**96.00**	3.08	97.41	**99.55**	98.44
	HIDRA3 (ours)	**3.30**	**0.176**	**4.26**	95.75	**1.02**	**98.23**	99.51	**98.82**
High SSH Values	$NEMO_0$	5.12	0.255	6.48	86.82	-2.96	92.20	**99.81**	95.24
	HIDRA3 (ours)	**4.46**	**0.245**	**6.04**	**89.94**	**-2.32**	**97.38**	98.81	**98.06**

Table 3. Performance of HIDRA3 and $NEMO_0$ under the target location tide gauge failure.

	Model	MAE (cm)	nMAE	RMSE (cm)	ACC (%)	Bias (cm)	Re (%)	Pr (%)	F1 (%)
Overall	NEMO	2.65	**0.142**	3.56	97.76	-0.31	/	/	/
	HIDRA2	2.63	0.146	3.56	98.15	-0.17	/	/	/
	$HIDRA3_1$ (ours)	**2.60**	0.144	**3.47**	**98.40**	**0.02**	/	/	/
Low SSH Values	NEMO	4.19	0.215	5.23	92.91	2.88	94.04	**99.92**	96.39
	HIDRA2	**3.27**	**0.175**	**4.27**	**95.94**	**1.02**	**97.64**	99.55	**98.51**
	$HIDRA3_1$ (ours)	3.52	0.190	4.47	95.58	1.79	97.31	99.21	98.16
High SSH Values	NEMO	4.68	0.244	6.19	89.14	-3.02	94.53	**99.40**	96.79
	HIDRA2	4.80	0.266	6.53	89.49	-2.35	96.62	97.82	97.18
	$HIDRA3_1$ (ours)	**4.34**	**0.239**	**5.98**	**90.94**	**-1.72**	**96.82**	98.54	**97.65**

Table 4. Performance of NEMO, HIDRA2 and $HIDRA3_1$, where $HIDRA3_1$ is the model trained separately on every single location.

Different sample sizes were used to calculate the error statistics, and we agree with the reviewer that this diminishes the validity of comparing the errors between stations. However, we prefer not to calculate the errors only at the time points where SSH data is available for all locations, as this would decrease the number of samples used in computing metrics at certain locations, undermining the validity of model comparisons at those sites. Fortunately, nearby stations tend to have similar data availability (as detailed below), so we have decided to retain the analysis in the manuscript as it is.

SSH data availability in the test period:

```
Koper
 2019: 100.0%
 2020: 100.0%

Venice
 2019: 99.7%
 2020: 93.0%


Ravenna
 2019: 0.0%
 2020: 97.1%

Ancona
 2019: 96.7%
 2020: 99.8%


Ortona
 2019: 0.0%
 2020: 99.1%

Tremiti
 2019: 0.0%
 2020: 92.8%

Vieste
 2019: 0.0%
 2020: 99.2%


Neretva
 2019: 99.9%
 2020: 100.0%

Sobra
 2019: 100.0%
 2020: 99.9%

Stari Grad
 2019: 99.9%
 2020: 99.4%

Vela Luka
 2019: 91.3%
 2020: 99.8%
```

COMMENT 4:

2.3. In Section 2.1 (line 84), the period from January 2019 to June 2019 is omitted from both training and testing. Is there a specific reason for this gap?

RESPONSE 4:

At the time of development CMEMS NEMO forecasting products were available only after June 2019, which is why we do not include the first part of 2019 in our evaluation. However, we should use that data for training to extend the training period.