Dear editor, dear reviewer.

Before we proceed with point-by-point addressing of the issues raised by the reviewer, we would like to make a very short synopsis of what was done during this revision.

1. We have found a minor date-matching error in our preprocessing scripts that had to be corrected. This required a full re-training of all HIDRA2 versions, including ablations and HIDRA1. While the numbers in Tables 1 and 2 are slightly altered due to retraining, the relative performance of HIDRA2 is negligibly altered by this change and all conclusions regarding performance from the original manuscript still hold.
2. During the code review, we noticed that F1, precision, and recall scores are not correctly calculated due to an error at the boundaries of each prediction and we corrected the code. The correction does not change the conclusions regarding performance made in the original manuscript. We have updated and published the code for the calculation of the scores.
3. We have thoroughly rewritten and expanded the section of the ablation study.

We proceed to a detailed point-by-point response below.

This manuscript shows an implementation of a neural network (HYDRA2) to predict sea level at the Koper tidal station. This neural network is an improvement over HIDRA1 and it is compared to its predecessor and to a numerical ocean model NEMO. Particularly notable in this manuscript is the detailed validation of the performance of the model. I can recommend publishing this manuscript after minor changes.

We thank the reviewer for encouraging comments.

To facilitate the reading of the manuscript and the interpretation of the figures and table I would recommend the captions clarify if the authors show an independent validation (data not used during training and not used for the optimization of hyperparameters, if this is the case) or validation with dependent data. Likewise I think it would be useful to mention this also with the skill scores mentioned in the abstract (starting at line 7) whether these error reductions are obtained from the independent test data or not.

We thank the reviewer for this remark. All validation is performed on an independent dataset that was hidden from the network during training. We now emphasize this fact throughout the revised manuscript to ensure clarity.

I don't not have any doubts about the scientific soundness of the results, but adding this information would help readers understand the results of the manuscript more quickly.

Minor comments:

Line 5: "single member of ECMWF atmospheric ensemble": is this the central forecast or any single member (chosen at random)?

We are using a fixed 42nd member of the atmospheric ensemble. Number 42 was chosen randomly to the extent that it is a tribute to the ultimate answer from the Hitchhiker's Guide to the Galaxy. Of course, over multi-year time intervals, this member is statistically completely equivalent to any other randomly selected member of the ECMWF ensemble prediction system. In other words, we could use any other ensemble member - or even choose a new random member in each run without substantially affecting the results. We now state this in the manuscript.

Line 47: "HIDRA1 ensemble (Žust et al., 2021) is a million times faster than the operational numerical ocean model ensemble based on NEMO engine (Madec, 2016) at Slovenian Environment Agency": There is not a lot of context to understand this comparison. NEMO will provide you with a sea level estimate over the whole domain. Is this also the case for HIDRA1 or would it provide the sea-level for a single location?

We thank the reviewer for this remark. Other reviewers raised the same issue and we have now amended the manuscript to better contextualize this comparison.

Our own ensemble setup of NEMO3.6 (used in our previous GMD paper on HIDRA1 but not in the present paper) is numerically expensive and time-consuming, so the forecasters and civil rescue obtain their daily forecasts shortly before noon each day. NEMO morning bulletins are thus mostly issued based on a NEMO ensemble run from the previous evening. HIDRA architectures on the other hand allow instantaneous forecast production (for a single point, Koper) as soon as we get ensemble and tide gauge data. This is an immediate benefit for the forecasting service and for civil rescue response.

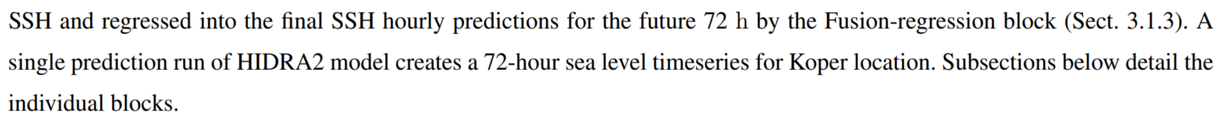
It is true that HIDRA computes predictions for a single point, while NEMO computes the sea state evolution of the entire basin, but this does not change the fact that HIDRA forecast of sea level is timely and NEMO's forecast generally is not. Warning triggers need only a single point sea level prediction, which HIDRA provides very fast and our in-house NEMO setup doesn't.

Line 99: "HIDRA2 does not require explicit annotation of whether a location point belongs to land or sea, thus land masks are not generated." I am wondering if the land-sea mask would still be a useful feature to provide to the neural network as a wind over land would not generate seiches. I guess that the neural network compensates for this by learning the land-sea mask internally.

We agree with the reviewer. Since the land mask is static, it is very likely that it is implicitly learned by the neural network, thus the network might not benefit from providing it at the input.

Line 103: "three-days prediction lead time" I think that your ML model will give in one application the full 3-day time series. Can you confirm? Or do you rather need to apply the ML model iteratively to obtain the 3-day time series? Can you also clarify this in the manuscript?

The reviewer is correct, HIDRA2 creates a 72-hour forecast in a single execution. We have made this clearer in the manuscript by adding the sentence that "A single prediction run of HIDRA2 model creates a 72-hour sea level timeseries for Koper location."



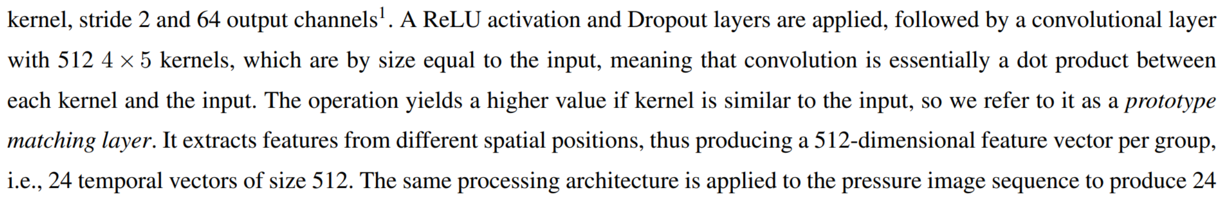
SSH and regressed into the final SSH hourly predictions for the future 72 h by the Fusion-regression block (Sect. 3.1.3). A single prediction run of HIDRA2 model creates a 72-hour sea level timeseries for Koper location. Subsections below detail the individual blocks.

104: "full ECMWF three-day forecast" -> "full" refers to the full ensemble (i.e. all ensemble members)?

Yes, that is correct. We have now included "(i.e. all ensemble members)" behind "full".

143: “prototype matching layer” Can you provide more information and a reference ?

Prototype matching refers to a convolution between the convolutional kernel pattern and the encoded input data. Convolution is an application of a dot product between the two at each location. The dot product at any location will be large if the feature at that location is similar to the convolution kernel and low otherwise. In this context, the learned convolutional kernels can be considered prototype patterns and the convolution operation as prototype matching. We have updated the manuscript to make the term clearer:

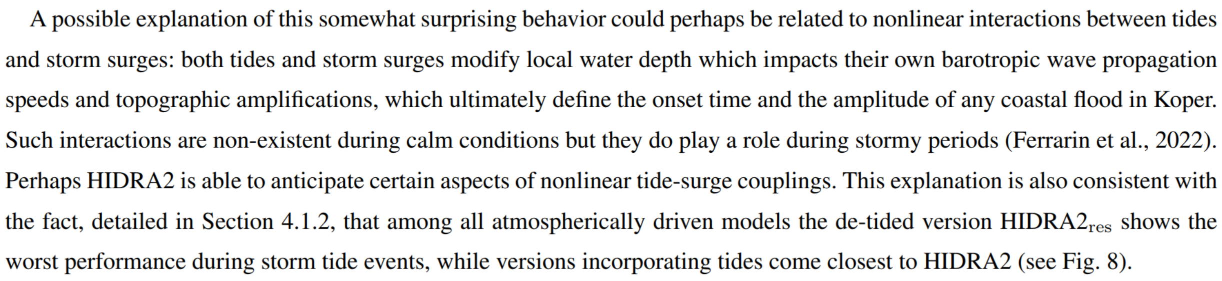


kernel, stride 2 and 64 output channels[1]. A ReLU activation and Dropout layers are applied, followed by a convolutional layer with 512 4×5 kernels, which are by size equal to the input, meaning that convolution is essentially a dot product between each kernel and the input. The operation yields a higher value if kernel is similar to the input, so we refer to it as a *prototype matching layer*. It extracts features from different spatial positions, thus producing a 512-dimensional feature vector per group, i.e., 24 temporal vectors of size 512. The same processing architecture is applied to the pressure image sequence to produce 24

section 4.1.1: this is an interesting and surprising result. Can the authors speculate why this is the case? (predicting full SSH leads to better results for extreme events). Could it be that the neural network internally limits its output range when working on anomalies? Do you expect this outcome to remain should one have more training data?

We thank the reviewer for this question. Interpretability of neural networks is an open research problem, thus we can only speculate why HIDRA2 benefits from total sea levels but not so much from the residuals.

It seems that HIDRA2 learns to extract some information from the full sea level signal which consists of linear and nonlinear interactions between the tides and the storm surge. One of such nonlinear interactions, for example, reflects the fact that both storm surge and the tides modify local undisturbed water depths which impact their own barotropic propagation speeds and their respective topographic amplifications. Perhaps HIDRA2 is capable of resolving certain aspects of such interplays of phenomena. This interaction is practically non-existent during calm periods and is also less pronounced in the detided residual signal. This might explain why the benefit of full SSH is most obvious during storm tides. We have added the following discussion to the manuscript:



A possible explanation of this somewhat surprising behavior could perhaps be related to nonlinear interactions between tides and storm surges: both tides and storm surges modify local water depth which impacts their own barotropic wave propagation speeds and topographic amplifications, which ultimately define the onset time and the amplitude of any coastal flood in Koper. Such interactions are non-existent during calm conditions but they do play a role during stormy periods (Ferrarin et al., 2022). Perhaps HIDRA2 is able to anticipate certain aspects of nonlinear tide-surge couplings. This explanation is also consistent with the fact, detailed in Section 4.1.2, that among all atmospherically driven models the de-tided version HIDRA2$_{\mathrm{res}}$ shows the worst performance during storm tide events, while versions incorporating tides come closest to HIDRA2 (see Fig. 8).

We are guessing that some further role is played by the fact that residual sea levels are contaminated with remnants of tidal signal but this contamination occurs in a very non-obvious way which is not related to the basin dynamics, but rather to our tidal algorithms, input data and (sea level - tide) subtraction.

Section 4.1.2 “Ablation study”: can you clarify that you retrained the network for the different test cases (without tides encoder, without atmospheric encoder,..) or you do rather zero-out the output of the corresponding encoder without re-training.

Thank you for pointing this out: indeed we do leave out different input streams and/or encoders every time and we retrain a different network every time. Ablation study therefore consists of training and evaluating different architectures. It is not a post-processing activity.

We have now thoroughly rewritten, expanded, and restructured the Ablation section to make our actions more transparent.

Typo:

Line 205: 1°/24 -> 1/24 °

We changed this to (1/24)°.