



HIDRA2: deep-learning ensemble sea level and storm tide forecasting in the presence of seiches – the case of the northern Adriatic

Marko Rus¹, Anja Fettich², Matej Kristan^{1,★}, and Matjaž Ličer^{2,3,★}

¹Faculty of Computer and Information Science, Visual Cognitive Systems Lab, University of Ljubljana, Ljubljana, Slovenia

²Slovenian Environment Agency, Office for Meteorology, Hydrology and Oceanography, Ljubljana, Slovenia

³National Institute of Biology, Marine Biology Station, Piran, Slovenia

★These authors contributed equally to this work.

Correspondence: Marko Rus (marko.rus@fri.uni-lj.si)

Received: 8 July 2022 – Discussion started: 29 August 2022

Revised: 14 November 2022 – Accepted: 3 December 2022 – Published: 10 January 2023

Abstract. We propose a new deep-learning architecture HIDRA2 for sea level and storm tide modeling, which is extremely fast to train and apply and outperforms both our previous network design HIDRA1 and two state-of-the-art numerical ocean models (a NEMO engine with sea level data assimilation and a SCHISM ocean modeling system), over all sea level bins and all forecast lead times. The architecture of HIDRA2 employs novel atmospheric, tidal and sea surface height (SSH) feature encoders as well as a novel feature fusion and SSH regression block. HIDRA2 was trained on surface wind and pressure fields from a single member of the European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric ensemble and on Koper tide gauge observations. An extensive ablation study was performed to estimate the individual importance of input encoders and data streams. Compared to HIDRA1, the overall mean absolute forecast error is reduced by 13 %, while in storm events it is lower by an even larger margin of 25 %. Consistent superior performance over HIDRA1 as well as over general circulation models is observed in both tails of the sea level distribution: low tail forecasting is relevant for marine traffic scheduling to ports of the northern Adriatic, while high tail accuracy helps coastal flood response. Power spectrum analysis indicates that HIDRA2 most accurately represents the energy density peak centered on the ground state sea surface eigenmode (seiche) and comes a close second to SCHISM in the energy band of the first excited eigenmode. To assign model errors to specific frequency bands covering di-

urnal and semi-diurnal tides and the two lowest basin seiches, spectral decomposition of sea levels during several historic storms is performed. HIDRA2 accurately predicts amplitudes and temporal phases of the Adriatic basin seiches, which is an important forecasting benefit due to the high sensitivity of the Adriatic storm tide level to the temporal lag between peak tide and peak seiche.

1 Introduction

Global mean sea level rise, related to anthropogenic climate change (Arias et al., 2021), is causing a worldwide increase in coastal flooding frequency (Taherkhani et al., 2020) and is leading to a myriad of negative consequences for coastal communities, civil safety and economies (Ferrarin et al., 2020). Shallow semi-enclosed coastal regional basins like the northern Adriatic (northern–central Mediterranean Sea) are facing growing threats of coastal inundation and erosion (Ferrarin et al., 2020), seawater intrusions into freshwater reservoirs and worsening conditions for marine traffic. Northern Adriatic ports like Venice, Koper and Trieste, but also other cultural landmark towns like Chioggia or Piran, have been – or will be – forced to take expensive preventive measures to mitigate their exposure.

The problem of sea level forecasting on the northern Adriatic shelf (see Fig. 1 for the shelf location and depth) is two-fold: (i) high sea levels lead to severe flooding of densely

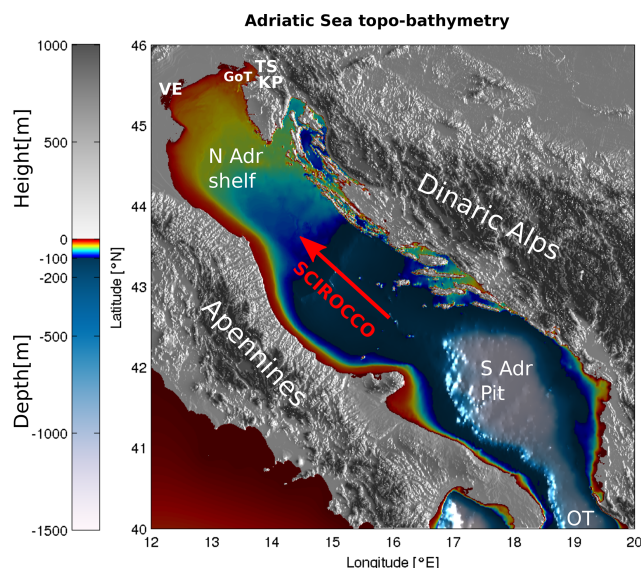


Figure 1. Topography and bathymetry of the Adriatic region. Abbreviations used on the map are as follows: TS – Trieste, KP – Koper, GoT – Gulf of Trieste, VE – Venice, N Adr shelf – northern Adriatic shelf, S Adr Pit – southern Adriatic pit, OT – Otranto Strait. The direction of the scirocco is marked with the red arrow. The image was created by the authors based on EMODnet bathymetry data, available at <https://portal.emodnet-bathymetry.eu/> (last access: 8 June 2022), and the Copernicus European Digital Elevation Model, available at <https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1-0-and-derived-products/eu-dem-v1.0> (last access: 8 June 2022).

populated coastal towns, while (ii) low sea levels may effectively inhibit large marine cargo due to very shallow depths (often below 15 m) of marine waterways on the shelf and especially in the Gulf of Trieste. Reliable forecasting of both tails, high and low, of sea level distribution is therefore imperative for services like civil safety and cargo scheduling activities in local ports.

The two distribution tails, however, represent two dynamically separate problems. High sea levels always occur due to intense pressure lows and corresponding strong winds during cyclonic activity in the basin, while extremely low sea levels typically occur through a combination of prolonged periods of high atmospheric pressure and spring tides.

Equilibrium ocean response to slow changes in air pressure is captured by the inverse barometer effect, while the wind setup of the sea level occurs through the vertical momentum flux across the air–sea interface. The dominant winds in the Adriatic basin are the southeasterly scirocco, blowing along the major axis of the basin (see Fig. 1), and the northeasterly cross-basin bora. Strong scirocco events lead to severe storm surges, excitation of basin seiches (Bajo et al., 2019) and potentially severe flooding in the northern Adriatic. Adriatic along-basin seiches have eigenperiods of 21.5 h (fundamental eigenmode) and 10.9 h (first excited

eigenmode) (see, e.g., Medvedev et al., 2020) and decay on the timescale of days, mostly due to radiation losses through Otranto (Cerovecki et al., 1997).

In this paper we will adhere to the terminology proposed in Gregory et al. (2019): (i) the term *sea level* will denote the total time-varying local water depth at the tide gauge in Koper, (ii) the term *sea surface height* will be the height of the sea level above (or below) the reference ellipsoid, (iii) the term *storm surge* will denote the elevation or depression of the sea surface with respect to the predicted tide during a storm, and (iv) the term *storm tide* will be the sea surface height, elevated during a storm by a storm surge.

The key difficulty of sea level forecasting in the Adriatic basin arises from the high sensitivity of the total sea level to the phase lag between the gravitationally generated tides (independent of meteorological forcing) and meteorologically generated basin seiches (independent of gravitational forcing). This sensitivity can translate reliable atmospheric forecasts with very limited errors in the timing and trajectory of the cyclone into substantial errors in the sea level forecast. Probabilistic ensemble forecasting of sea level envelopes with error variance estimation (Žust et al., 2021; Ferrarin et al., 2020; Bernier and Thompson, 2015; Mel and Lionello, 2014) was therefore explored to tackle this drawback. However, ensemble sea level forecasting is numerically expensive, requires specialized expensive hardware, and introduces delays in prediction. To avoid the high numerical cost of ensemble sea level forecasting, computationally efficient machine-learning-based ensemble models have recently been explored (Žust et al., 2021). While these models require a substantial amount of training data to learn the complex relations for reliable predictions, the inference is numerically cheap. For example, single-point Koper sea level predictions from the neural network HIDRA1 ensemble (Žust et al., 2021) are a million times faster than the full-basin operational NEMO ocean (Madec, 2016) at the Slovenian Environment Agency. It is true that HIDRA1 computes prediction for a single variable in a single point, while ocean models compute the 4D evolution of a broad set of oceanic properties, but in the operational environment, faster model prediction times come with immediate benefits for downstream warning issuing and civil rescue operations.

Machine learning has thus been explored by several research groups for single-point sea level forecasting. The early approaches (Imani et al., 2018) were based on classic machine learning models such as support vector machines (Sapankevych and Sankar, 2009) with radial basis function kernels. In their work, Pashova and Popova (2011) and Karimi et al. (2013) utilized shallow fully connected neural networks, but due to simplistic network architectures that did not utilize the numerical atmospheric forecast, they could only report the desired accuracy for short temporal horizons. Ishida et al. (2020) used long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) together with several atmospheric variables to improve 1 h pre-

diction into the future but did not expand the prediction horizon. Braakmann-Folgmann et al. (2017) predicted further in time by applying a combination of LSTM and convolutional neural networks but at a very coarse level. Autoregressive neural networks were considered in Hieronymus et al. (2019) to increase the temporal resolution and the prediction horizon. Most recently, a convolutional neural network, HIDRA1 (Žust et al., 2021), with a specialized architecture to utilize atmospheric data, sea surface heights and astronomic tides was proposed. To the best of our knowledge, HIDRA1 is currently the most accurate machine learning sea surface height prediction model, with a prediction horizon of several days in length at an hourly resolution. However, while HIDRA1 performed favorably in comparison to the NEMO model used in that study, it failed to beat the NEMO setup at very high and very low ends of sea level distributions. In other words, extreme sea levels (coastal floods on the one end and very low sea levels on the other), which interest us the most, were not yet captured with sufficient reliability and present an open challenge for machine learning methods.

In this paper we propose HIDRA2, our latest attempt at sea level forecasting using deep learning. In contrast to the previous version, HIDRA2 presents a novel architecture with new atmospheric, tidal and sea surface height (SSH) feature encoders as well as a novel feature fusion and SSH regression block. An additional conceptual novelty is that HIDRA2 predicts the full SSH rather than the residual (i.e., the difference between SSH and astronomic tide), as is the case for HIDRA1. The new model extracts relevant information from different spatial locations in the atmosphere signal and predicts the SSH with a 3 d horizon at an unprecedented accuracy, outperforming HIDRA1 as well as two state-of-the-art ocean models.

The paper is organized as follows. Section 2 introduces the sea level and atmospheric model data and performance measures used in our analysis. Section 3 details the new HIDRA2 architecture and the numerical ocean model setup used as the performance baseline. Section 4 reports the analysis of the HIDRA2 architecture (including an extensive ablation study) and provides detailed quantitative as well as qualitative comparisons with the state-of-the-art numerical ocean models. Conclusions and outlook are given in Sect. 5.

2 Training and evaluation datasets

2.1 Sea level training data

SSH observations during the period 2006–2018 were retrieved from Koper Mareographic Station (45°33'N, 13°44'E; see Fig. 2 for the location), which is maintained by the Slovenian Environment Agency (ARSO) and is part of the European Sea Level Network (Pérez Gómez et al., 2022). Observations are obtained in 10 min intervals using both a float-type sensor and an additional radar sea level sen-

sor, and they undergo subsequent quality control at ARSO (Pérez Gómez et al., 2022). Hourly data points are extracted to get the signal used in HIDRA2 training and evaluation.

The tidal signal in the sea level is independent of atmospheric processes and can be computed by tidal analysis and prediction models. The tidal contribution to Koper SSH considered in this study is estimated from hourly instantaneous SSH values in 1-year segments using the UTIDE Tidal Analysis package for Python (Codiga, 2011). The total sea level series is then decomposed into a tidal part and a residual part, where we define the sea level residual as the arithmetic difference between the total sea level and the tidal sea level. According to the ARSO operational protocol, the SSH is classified as a flood if it is higher than 300 cm. Floods thus constitute 0.41 % of all training data.

2.2 Atmospheric training data

Atmospheric input used for HIDRA2 training was retrieved from the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (Leutbecher and Palmer, 2007). The ECMWF ensemble forecasts come as a global atmospheric ensemble of 50 members with a 0.125 arc degree spatial resolution and a 3 h temporal resolution. The training dataset used in this study consists of (i) 10 m winds and (ii) mean sea level air pressure from a single fixed (42nd) atmospheric ensemble member during the period 2006–2018. 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* (Adams, 1979). Of course, over multiyear time intervals, this member is completely statistically equivalent to random use of any other member of the ECMWF ensemble prediction system. In other words, we could use any other ensemble member – or choose a different random member each run – without substantially affecting the results. All ECMWF input fields were standardized and cropped to the Adriatic basin, represented by a 57×73 spatial grid (see Fig. 2). The forecasts were linearly interpolated to hourly time steps to match the SSH temporal resolution. To simplify the training protocol, a single atmospheric sequence is constructed by concatenating the first 24 h of daily consecutive ECMWF forecasts into the final atmospheric sequence used in training. HIDRA2 does not require explicit annotation of whether a location point belongs to land or sea, and thus land masks are not generated.

2.3 Evaluation data

The evaluation input dataset for both HIDRAs and NEMO is disjoint from the training dataset (years 2006–2018) and consists of ECMWF atmospheric predictions and Koper sea levels between 1 June 2019 and 31 December 2020. This period was chosen due to challenging conditions and an unusually high incidence of floods. We use the ECMWF daily predictions, each containing 50 ensemble members with 3 d

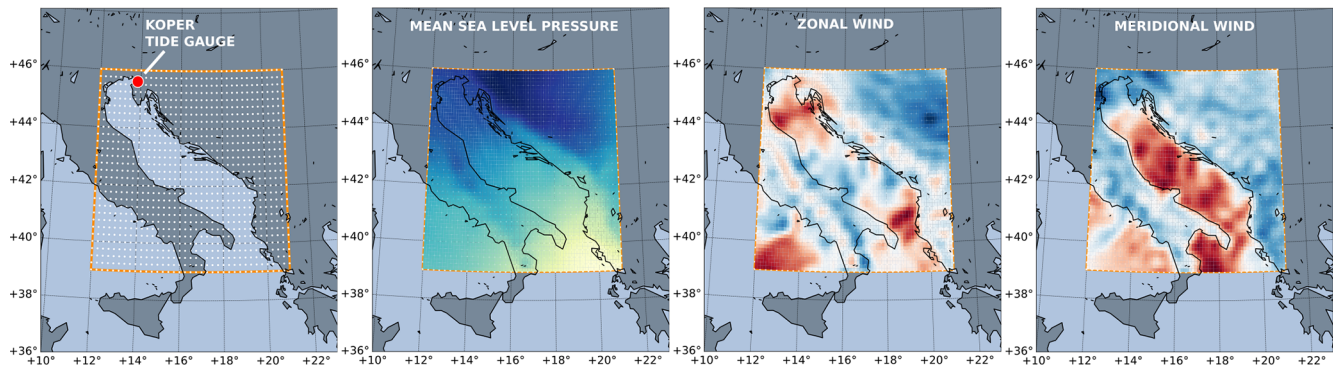


Figure 2. HIDRA2 input domain and dataset. The leftmost panel depicts the ECMWF grid (white dots) and Koper tide gauge location (red circle). Three panels on the right depict snapshots of ECMWF atmospheric fields used during training.

prediction lead time. The data are standardized, and the dimensionality of the atmospheric data is reduced in the same fashion as described in Sect. 2.2, except that for inference, the full (i.e., containing all ensemble members) ECMWF 3 d forecast is presented to the model. The floods represent 1.1 % of the test dataset.

2.4 Performance measures

Standard measures, i.e., the mean absolute error (MAE), the root mean square error (RMSE) and the model bias, are used to evaluate prediction performance. To reflect the practical suitability, we additionally calculate the prediction accuracy as a ratio between the predictions which are within 10 cm of the ground truth and all predictions. This 10 cm threshold reflects an acceptable deviation from the ground truth and was determined through discussion with the operational forecasting service at ARSO. The metrics are calculated globally by considering all prediction points as well as separately only on floods to reflect the prediction performance at these critical rare events.

To further probe the flood event prediction capabilities, we make use of the standard performance measures from the detection literature: precision Pr , recall Re and the F1 measure $F1$. Firstly, we need to define the flood event and then define the notion of the event being detected. Both of these have been defined in discussion with operational forecasters at ARSO. The anchor (i.e., temporal point) of a flood event is defined as the time of the local maximum in an SSH sequence above 300 cm. If the predicted flood event anchor is within a 3 h margin (before or after) from the nearest ground truth flood event anchor, it is considered a true positive TP; otherwise, it is a false positive FP. A flood event in the ground truth is considered a false negative FN if there is no matching flood event anchor in the predicted SSH. Like in the accuracy definition, the tolerance of 10 cm is applied, meaning that predictions below 300 cm are also considered to be TPs when they appear within the margin of 10 cm and that false positives with ground truth within 10 cm are ignored. The

precision and recall are then calculated as

$$Pr = \frac{TP}{TP + FP}, \quad Re = \frac{TP}{TP + FN}, \quad (1)$$

while the F1 measure that summarizes the detection performance, i.e.,

$$F1 = 2 \frac{Pr \cdot Re}{Pr + Re}, \quad (2)$$

is defined as the harmonic mean between precision and recall.

3 Numerical models

3.1 HIDRA2

The proposed HIDRA2 is the second generation of a deep neural model for sea surface height prediction, with HIDRA1 (Žust et al., 2021) being the first. The new architecture is shown in Fig. 3. The input data are encoded by three encoders: the wind and pressure sequences for the past 72 h are processed and merged by the Atmospheric encoder (Sect. 3.1.1), the tidal signal for the future 72 h is encoded by the Tidal encoder, and the sea surface height measurements coupled with the tide for the past 24 h are encoded by the SSH encoder (Sect. 3.1.2). The outputs of all three encoders are re-calibrated, fused with the past 72 h 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 the HIDRA2 model creates a 72 h sea level time series for the Koper location. The subsections below detail the individual blocks.

3.1.1 Atmospheric encoder

The atmospheric data for the Adriatic basin at a given time step are represented by a 57×73 spatial grid, i.e., an image. HIDRA2 assumes that the coarse spatial resolution of the atmospheric data contains enough information to provide

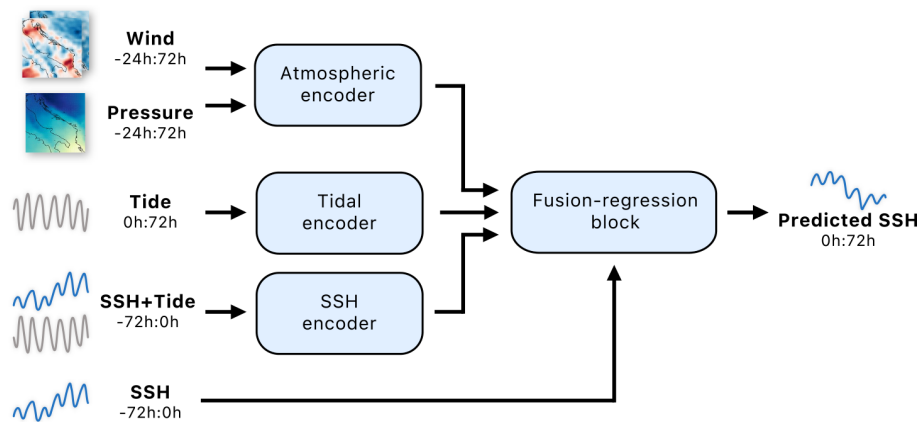


Figure 3. The HIDRA2 architecture. The Atmospheric encoder embeds the wind and pressure sequences with learnable temporal subsampling and pattern prototype matching to extract relevant features from different geographic locations and fuse them temporally into a single feature embedding. The Tidal and SSH encoders encode the future tide evolution and the past SSH and tide observations, respectively. All features are re-calibrated, fused with the past 72 h SSH, and regressed into the final SSH predictions by the fusion-regression block. Notation $a:b$ indicates hourly data points from the interval $(a, b]$, while the prediction point is at index 0.

satisfactory results, so it first downsamples the atmospheric data to a 9×12 grid by an average pooling operation.

The Atmospheric encoder is composed of two stages. In the first stage (shown in Fig. 4), the sequences of the wind and pressure images are independently processed by their respective encoding blocks, which use the same architecture. The wind image sequence of 96 h (the past 24 and future 72 h) is divided into 24 groups of 4 consecutive hours, which are processed independently. The spatial and temporal dimension of each group is reduced by a learnable 2D convolutional layer with a 3×3 kernel, stride 2 and 64 output channels¹. 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 the 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 512D 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 vectors of size 512. The two outputs are then concatenated to form a mixed set of 24 wind–pressure features of size 1024.

The second stage of the Atmospheric encoder (Fig. 5) extracts the temporal atmospheric features by considering the consecutive wind–pressure features extracted by the first stage. A 1D convolutional layer with kernel temporal dimension size 5 and with 256 output channels² is applied, entangling the information from temporal segments equivalent to

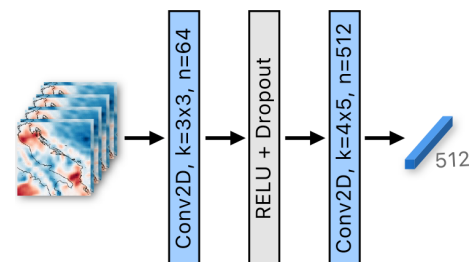


Figure 4. The first stage of the Atmospheric encoder. The input are 4 consecutive hours with two wind channels (this case) or pressure. 3×3 convolution is applied, followed by a prototype matching layer, outputting a single vector of size 512. Note that 24 independent passes are performed in parallel for the entire atmospheric input sequence. The variables k and n denote the kernel size and the number of output channels, respectively.

20 h in length. Note that because we are using a convolutional layer instead of the fully connected layer, the number of learnable parameters of the entire Atmospheric encoder is independent of the forecast horizon. Each of the obtained 20 features³ is then independently processed by a network containing two blocks of residual connections, each involving 1D convolution with kernel temporal dimension size 1 (i.e., 1×256 kernels), a SELU activation (Klambauer et al., 2017) and a dropout layer. Finally, each of the obtained 20 256D features are convolved by 32 1×256 kernels to reduce their dimensionality to 20×32 .

¹Note that the number of output channels is equal to the number of different kernels used in the layer.

²Note that the sizes of the kernels in the 1D convolutional layer are 1024×5 , but with the first dimension matching the size of the input features, the convolution displacements are only along the sec-

ond dimension, hence the 1D convolutional direction implied by the layer's name.

³The fact that the number of output segments is equal to the 20 h timespan is coincidental.

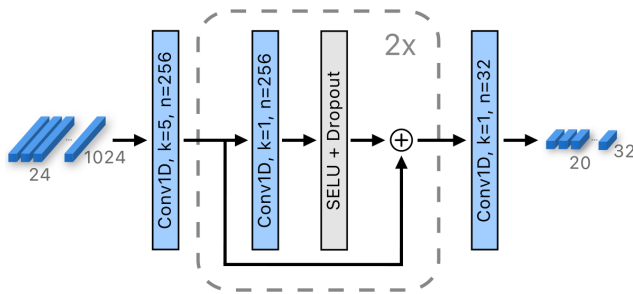


Figure 5. The second stage of the Atmospheric encoder. Features from all time points and both wind and pressure are processed with a 1D convolution, followed by two blocks with residual connections. The last convolution reduces feature dimensionality. The variables k and n denote the kernel size and the number of output channels, respectively.

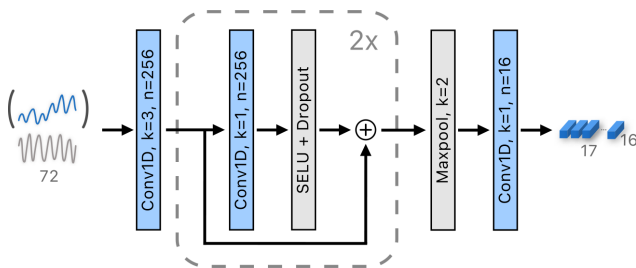


Figure 6. The SSH encoder encodes a concatenation of the past SSH and tide by a 1D convolution, followed by two blocks with residual connections, max-pooling temporal reduction and convolution-based feature reduction. The variables k and n denote the kernel size and the number of output channels, respectively.

3.1.2 Tidal and SSH encoders

Both the tidal and SSH encoders use the same architecture, the only difference being the size of the encoders' input. Figure 6 depicts the SSH encoder, which takes as input the past 72 h of SSH measurements and the tide concatenated into a 72×2 input tensor and processes them in a similar fashion to the second stage of the Atmospheric encoder: the input is processed by convolution with 256×2 kernels, which is followed by two consecutive residual blocks, a subsampling max-pooling layer and a final convolutional layer to reduce the dimensionality of the features to 17×16 . The Tidal encoder follows the same architecture, the only difference being that the input is the tide forecast for the next 72 h.

3.1.3 Fusion-regression block

The Atmospheric, Tide and SSH encoders produce temporal features of different importance and size. To account for that, the features are re-calibrated by normalization with means and variances of the features calculated during training and then denormalized with learned weights and biases. The form of normalization follows the batch normalization layer (Ioffe

and Szegedy, 2015), which applies a 0.9 momentum for updating the running means and variances during training. The normalized features are then concatenated and mixed by a fully connected layer, reducing their final dimension from 1184 to 512 (left part of Fig. 7). The obtained 512D domain context feature vector thus contains rich atmospheric and SSH information from all time points and all parts of the input domain.

While the encoding and mixing operations extract the domain context, the explicit surface height information might not be well retained in the extracted feature vector. To re-inject this information, the obtained domain context feature vector is concatenated with the time series of past observed SSH before passing to the final regression block. The latter is composed of two fully connected layers with 584 units, SELU activations and residual connections, followed by a fully connected layer with 72 outputs for the 72 h prediction horizon (see Fig. 7).

3.1.4 The network training

HIDRA2 is trained end to end using mean squared error (MSE) loss between the predictions and the ground truth. We train the model using the AdamW optimizer (Loshchilov and Hutter, 2017) with standard parameter values (learning rate 1×10^{-4} and running average damping parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$) and apply the cosine annealing (Loshchilov and Hutter, 2016) learning schedule to gradually reduce the learning rate during training to 1×10^{-5} . The training batch size is set to 512 data samples, and the model is trained for 40 epochs. Training takes approximately 1.5 h on a single computer with an NVIDIA GeForce RTX 2080 TI graphics card, while the inference of a single 72 h prediction for one member of the atmospheric ensemble takes only 4 ms.

3.1.5 Summary of differences to HIDRA1

While there are many differences between HIDRA2 and HIDRA1, we summarize only the major conceptual ones for a clearer exposition of the contributions. HIDRA1 uses wind, pressure and 2 m temperature from ECMWF predictions, while our preliminary study showed that the new HIDRA2 architecture does not benefit from the temperature, and thus only wind and pressure are considered. HIDRA1 concatenates all atmospheric inputs at a time step and encodes them by Resnet (He et al., 2016) blocks. While Resnet excels in computer vision tasks that rely on high-level semantic feature abstraction, we argue that tailored shallower encoders are more appropriate for the extraction of meaningful atmospheric patterns. HIDRA2 thus separately encodes the wind and pressure by shallow encoders, which apply spatial pattern feature extraction, and then mixes the features from the two atmospheric variables by extracting temporal patterns. While this allows HIDRA2 to extract multiple spatial patterns in the data, only a single set of spatial weights is

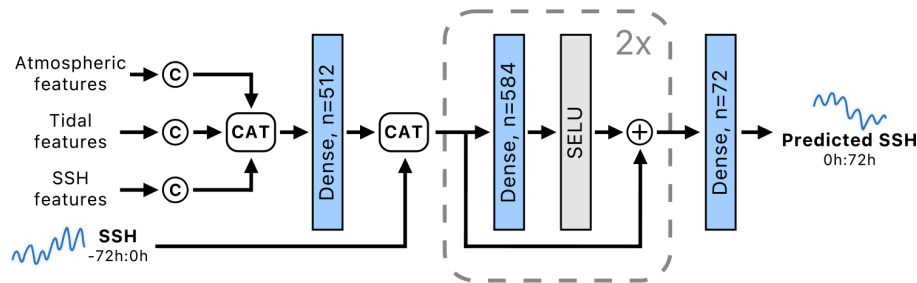


Figure 7. The fusion-regression block firstly re-calibrates the features (the C symbol), and then concatenated features are passed to a dense layer, which fuses features and reduces their dimensionality. Undistorted SSH is appended and processed with two residual blocks. The final dense layer outputs the predictions. The variable n denotes the number of output channels.

used to fuse the atmospheric features at a given time step in HIDRA1, consequently reducing its expressive power. HIDRA1 first averages 4 h atmospheric input data to temporally subsample the input, while HIDRA2 considers per-hour inputs and learns the appropriate spatiotemporal subsampling to maximize its predictive power. Another advantage of HIDRA2 is that it encodes the SSH input and mixes it with the atmospheric features early in the network to create a domain context feature vector before the final regression, while HIDRA1 considers only the atmospheric data for the context vector. Finally, HIDRA1 predicts the SSH residual (i.e., the difference between SSH and the astronomic tide), while HIDRA2 directly predicts the full SSH.

3.2 Ocean models

In this section we briefly describe two different numerical ocean modeling setups used for benchmarking HIDRA2. The two setups differ in several important respects. One is based on the NEMO ocean engine (Madec, 2016) and the other on the SCHISM (Zhang et al., 2016) modeling environment. The NEMO setup is described in more detail in Sect. 3.2.1, and it is a forecasting setup. The SCHISM setup is described in Sect. 3.2.2, and it is a reanalysis setup (Toomey et al., 2022). For brevity we will refer to the two setups presented below simply as NEMO or SCHISM.

3.2.1 NEMO ocean model

The Copernicus Marine Environment Monitoring Service (CMEMS) product MED-SEA_ANALYSISFORECAST_PHY_006_013 (see Clementi et al., 2021) was used as one of two numerical baselines for HIDRA2. This product is based on a Mediterranean basin configuration of the NEMO ocean model (Madec, 2016) and provides daily ocean forecasts for sea surface height above the geoid, temperature, salinity, circulation and mixed layer depth. The model domain spans the entire Mediterranean basin with a $(1/24)^\circ$ resolution and has 141 unevenly spaced vertical levels. The model solutions are operationally constrained to near-real-time observations

using a 3D variational assimilation scheme of temperature, salinity and along-track satellite sea level anomaly observations. The atmospheric forcing to the CMEMS model is provided by the ECMWF. Further details about the modeling setup can be found in Clementi et al. (2021). In this study, an SSH time series at the closest point to the Koper tide gauge was extracted from the Mediterranean ocean model forecast.

3.2.2 SCHISM ocean model

A barotropic setup of the SCHISM storm surge and wind-wave modeling environment (Toomey et al., 2022) was used as a second numerical baseline for HIDRA2. In this study, a single SSH time series from SCHISM reanalysis (Toomey et al., 2022) at the closest point to the Koper tide gauge was extracted and used for comparisons to the HIDRA models. SCHISM runs on an unstructured mesh covering the entire Mediterranean basin and extending into the Atlantic Ocean in the west. Its lateral boundary is forced by an equilibrium-inverted barometer ocean response to atmospheric pressure, while its surface forcing consists of ERA5 surface fields. The SCHISM unstructured grid allows for very high coastal resolutions, reaching some 200 m close to the coast.

3.2.3 Ocean model offset adjustment

Both NEMO and SCHISM sea levels, denoted here jointly as y^{model} , at any given location reflect departures from the local geoid and hence do not represent the absolute local depth of the water. The latter is furthermore also driven by low-frequency processes on the scales of many weeks or months, which are often difficult to capture for regional basin models on synoptic timescales. Prior to benchmarking, model results therefore have to be offset-adjusted to obtain total sea levels (required by port authorities and civil rescue) as follows. A time-averaged model (NEMO or SCHISM) SSH offset ϵ_n on the n th hour of the forecast day is defined as

$$\epsilon_n = n^{-1} \sum_{k=1}^n [y^{\text{model}}(t_k) - y^{\text{kp}}(t_k)], \quad (3)$$

where $y^{\text{kp}}(t)$ is the observed Koper sea level.

Each day, the value of ϵ_{12} is subtracted from $y^{\text{model}}(t)$ to ensure a zero bias for the first 12 h of the model day. Note that, despite this adjustment, the complete 72 h modeled time series may still exhibit a non-zero bias. Similar offset adjustment is not required for the HIDRA2s, since they predict the full SSH and learn to appropriately adjust for the offset automatically.

4 Results and discussion

4.1 HIDRA2 architecture analysis

As noted in Sect. 2, HIDRA2 was trained on the period 2006–2018 and evaluated on the period between 1 June 2019 and 31 December 2020. For evaluation, a single prediction is obtained by averaging predictions over all 50 ECMWF ensemble members. In the following we analyze the architectural choices of HIDRA2. The prediction of full SSH is justified in Sect. 4.1.1, while Sect. 4.1.2 reports an ablation study that aims to determine the role of specific encoders and types of input data. The results of all experiments are collected in Table 1.

4.1.1 Predicting the full SSH vs. the residual

A valid hypothesis can be made that predicting the residual (i.e., the difference between the full SSH and the tide) might be more beneficial than predicting the full SSH, since the network parameters might be better utilized by focusing only on the part of SSH not affected by the astronomic tide. In fact, HIDRA1 (Žust et al., 2021) does exactly this – it accepts and forecasts the residual. To explore this hypothesis, HIDRA2 was modified to predict only the residual ($\text{HIDRA2}_{\text{res}}$) by replacing the SSH input in the SSH encoder with the residual as well as in the fusion-regression block. Results in Table 1 indicate a similar overall performance when only the residuals are considered in HIDRA2. However, considering only stormy periods, we observe a substantial increase in the prediction error (+17.4 % MAE). This means that full SSH prediction is very beneficial for predicting floods while incurring only a small drop in the overall performance.

A possible explanation for 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 negligible 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 Sect. 4.1.2, that among all atmospherically driven models the de-tided version $\text{HIDRA2}_{\text{res}}$ shows the worst per-

formance during storm tide events, while versions incorporating tides come closest to HIDRA2 (see Fig. 8).

4.1.2 Ablation study: the importance of encoders and input data

An ablation study was executed to evaluate the importance of individual encoders and input data types. To estimate encoder importance, we removed each of the encoders in a separate experiment (and withheld all of their input data; see Fig. 3) and retrained the thus obtained ablated network. Ablation training and evaluation were conducted on identical datasets as with HIDRA2: years 2006–2018 represented the training set, and the time window between 1 June 2019 and 31 December 2020 served as an independent validation set. Note that, regardless of the encoder input, HIDRA2 always receives unencoded SSH data directly into its fusion-regression block (bottom dataflow branch in Fig. 3).

The following encoder ablations were performed.

1. Removal of the Atmospheric encoder ($\text{HIDRA2}_{\text{atmE}}$). Network $\text{HIDRA2}_{\text{atmE}}$ obtained no atmospheric input data, but it did receive SSH and tidal data.
2. Removal of the Tidal encoder ($\text{HIDRA2}_{\text{tidE}}$). Network $\text{HIDRA2}_{\text{tidE}}$ obtained no tidal input data for tidal encoding, but it did receive SSH and tidal data through the SSH encoder and atmospheric data through the Atmospheric encoder.
3. Removal of the SSH encoder ($\text{HIDRA2}_{\text{sshE}}$). Network $\text{HIDRA2}_{\text{sshE}}$ received atmospheric and tidal data through the Atmospheric and Tidal encoders, but it did not receive any SSH input via the SSH encoder.

The results in Table 1 show that MAE increases with each modification, particularly during storm events. Removal of the Atmospheric encoder results in the most significant performance drop, indicating that the atmospheric features convey by far the most relevant predictive information. A significant performance drop is observed as well when removing the Tidal encoder. The SSH encoder has the smallest impact on overall performance yet still importantly contributes to the prediction accuracy during storms.

Two further ablations were then performed regarding the data types of the sea level input data (the SSH and the tide; see Fig. 3) which are considered in the SSH encoder. We retained HIDRA2 with all three of its encoders but provided the SSH encoder with limited sea level input.

1. Removal of the tidal input to the SSH encoder ($\text{HIDRA2}_{\text{tidI}}$). In this case the SSH encoder received as input only the total sea level.
2. Removal of the SSH input ($\text{HIDRA2}_{\text{sshI}}$). In this case the SSH encoder received as input only the tidal sea level.

Table 1. Performance of ablated HIDRA2 designs evaluated over all sea level bins (the *Overall* column) and only on storm tide events (*Storm tide events* column). The evaluation period spans 1 June 2019–31 December 2020, which is completely independent of the training data. Row (a) shows the performance of HIDRA2 in predicting the residual, row (b) shows the performances of encoder ablations, and row (c) corresponds to SSH input ablations and re-calibration. The final row (d) corresponds to the final version of HIDRA2. The best scores are formatted in bold.

Modification	Overall				Storm tide events						
	MAE (cm)	RMSE (cm)	Bias (cm)	Acc (%)	MAE (cm)	RMSE (cm)	Bias (cm)	Acc (%)	Re (%)	Pr (%)	F1 (%)
(a) HIDRA2 _{res}	4.11	5.84	−0.69	92.77	11.47	15.52	−9.13	57.20	76.16	91.27	83.03
(b) HIDRA2 _{atmE}	7.54	11.40	−0.53	75.75	27.86	34.28	−26.21	23.01	41.72	91.30	57.27
HIDRA2 _{tidE}	4.60	6.37	−0.02	90.66	10.57	14.82	−7.74	61.94	78.81	92.25	85.00
HIDRA2 _{sshE}	4.24	5.97	−0.37	92.45	10.80	15.01	−8.09	61.08	80.13	93.80	86.43
(c) HIDRA2 _{tidI}	4.24	5.96	−0.47	92.28	10.65	14.74	−7.63	58.92	80.13	93.08	86.12
HIDRA2 _{sshI}	4.23	5.93	−0.36	92.59	11.04	15.14	−7.99	58.49	82.78	91.24	86.81
HIDRA2 _{norm}	4.14	5.85	−0.03	92.72	10.38	14.35	−7.48	60.43	81.46	92.48	86.62
(d) HIDRA2	4.12	5.82	0.21	92.89	9.77	14.07	−5.99	64.52	84.11	91.37	87.59

The results in Table 1 show that the removal of each leads to a consistent but moderate increase in the errors overall. However, the errors increase substantially during storms, indicating the importance of using both types of inputs.

We observe a similar situation when removing the atmospheric and SSH/tide feature re-calibration in the fusion-regression block (HIDRA2_{norm}). Results in Table 1 indicate that feature normalization does not affect performance in normal conditions, but it substantially contributes to the prediction accuracy of storm tides. A closer inspection of HIDRA2_{norm} showed that the scale of the tidal features is 4 times larger than the scale of the atmospheric features. Inclusion of the re-calibration blocks, however, remedies this by making the scales of all features (atmospheric, SSH and tidal) approximately the same.

Figure 8 depicts performances of ablated HIDRA2 versions across all sea level bins. Even though most global performance metrics of HIDRA2 (depicted in Table 1) are the best, Fig. 8 indicates that, for low sea levels, HIDRA2_{res} exhibits slightly lower errors. HIDRA2, however, performs substantially better in the flooding regime above 300 cm. This further substantiates our final choice of HIDRA2 architecture.

4.2 Comparison with the state-of-the-art numerical ocean models

HIDRA2 is compared with HIDRA1 (Žust et al., 2021), which is currently the state of the art in machine learning SSH prediction (Sonnewald et al., 2021) and with state-of-the-art numerical ocean modeling setups NEMO (Madec, 2016) and SCHISM (Toomey et al., 2022). The methods are evaluated on an independent time window (1 June 2019–30 December 2020) and with respect to different SSH values (see Sect. 4.2.1), Sect. 4.2.2 reports performance with

respect to the lead times, while spectral analysis is reported in Sect. 4.2.3. The last two sections discuss performances on historical storm surge events (Sect. 4.2.4) and the forecast spectral decomposition of these events (Sect. 4.2.5).

4.2.1 SSH forecast performance

The overall prediction performance and the performance restricted to storm events are shown in Table 2. HIDRA2 outperforms HIDRA1, NEMO and SCHISM overall as well as during storms, yielding a lower MAE/RMSE and higher accuracy. While HIDRA1 achieves a lower bias, its RMSE and MAE are substantially higher – HIDRA2 outperforms HIDRA1 in MAE by 12.7 % overall and by 24.6 % during the storm tide events. NEMO achieves the highest precision of flood detection (Pr = 100 %), meaning that all detected floods are true positives. However, while all NEMO's predicted floods were true, not all floods were predicted, resulting in its low recall of Re = 63.58 %. A similar situation is observed for HIDRA1. The recalls for these two methods (NEMO: 63.58 % and HIDRA1: 74.17 %) are substantially lower than that of HIDRA2 (Re = 84.11 %), which detects many more floods with fewer false negatives. The excellent trade-off between the precision and recall of HIDRA2 is reflected in its F1 score (87.59 %), which is substantially higher than that of NEMO (77.73 %), HIDRA1 (82.96 %) or the next best SCHISM (83.80 %).

For detailed analysis, we visualize the MAE values of the tested methods with respect to the sea level heights in Fig. 9. HIDRA2 consistently shows the lowest errors at all sea level bins. During storm tides, NEMO outperforms HIDRA1, while HIDRA2 and SCHISM outperform both HIDRA1 and NEMO by several centimeters. Solid HIDRA2 performance at the low end of the sea level distribution is particularly important to note because of its potentially high significance to

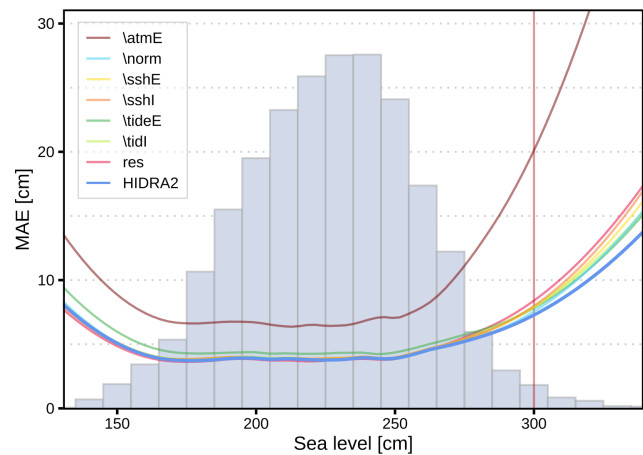


Figure 8. Mean absolute error (MAE) of ablated HIDRA2 designs evaluated over all sea level bins. Vertical red line indicates the flooding threshold in Piran. Performances of all the models were evaluated on a 1 June 2019–31 December 2020 dataset, which is completely independent of the training data.

Table 2. Performance of HIDRA1, HIDRA2, NEMO and SCHISM over all sea level bins (the *Overall* column) and only during storm tide events (*Storm tide events* column). Tidal forecast is included for reference. The evaluation period spans 1 June 2019–31 December 2020, which is completely independent of the training data. The best scores are formatted in bold.

	Overall				Storm tide events						
	MAE (cm)	RMSE (cm)	Bias (cm)	Acc (%)	MAE (cm)	RMSE (cm)	Bias (cm)	Acc (%)	Re (%)	Pr (%)	F1 (%)
Tide	13.82	18.86	−5.13	47.45	55.75	59.45	−55.75	0.00	0.00	/	/
NEMO	6.54	8.52	−1.23	79.14	13.03	17.09	−11.24	49.68	63.58	100.00	77.73
SCHISM	5.57	7.50	0.20	85.06	11.04	14.70	−6.19	57.63	78.81	89.47	83.80
HIDRA1	4.72	6.73	−0.26	90.04	12.95	17.65	−10.66	53.76	74.17	94.12	82.96
HIDRA2	4.12	5.82	0.21	92.89	9.77	14.07	−5.99	64.52	84.11	91.37	87.59

marine traffic scheduling in the very shallow seas surrounding the port of Koper, which is currently restricted to periods of high tides. In summary, HIDRA2 outperforms all state-of-the-art methods for all sea level heights, thus displaying a solid prediction skill in moderate as well as extreme values of the sea surface height.

4.2.2 Performance with regard to forecast lead time

We next analyzed how the prediction lead time affects the prediction errors. Figure 10 shows the MAE scores with respect to the prediction lead time for the values between 1 and 72 h. The MAE of the prediction gradually increases with the lead time for all the tested methods. While overall a solid performer with a MAE well below 10 cm, NEMO exhibits the highest MAE and also the highest MAE variance. Clear signals are observed with 12 and 24 h periods in the NEMO MAE. Since NEMO includes tides, we suspect this periodicity stems mostly from the errors in either amplitude or phase of the tidal part of the NEMO sea level signal, but further research would be necessary to properly substantiate this claim. SCHISM shows better performance (lower MAE)

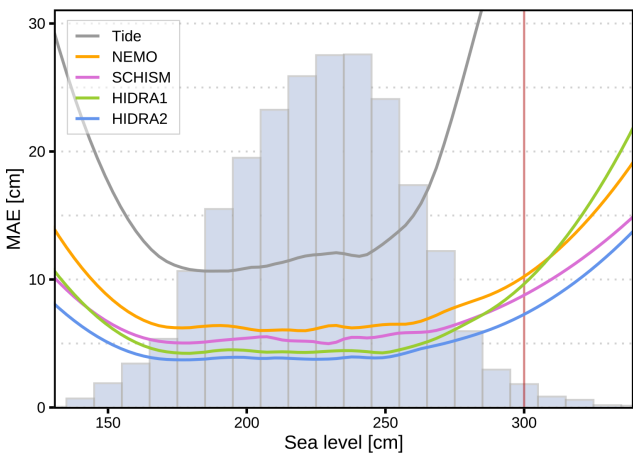


Figure 9. HIDRA, NEMO and SCHISM performances with regard to sea level bins (grey histogram in the bottom layer). The coastal flood threshold is marked with a vertical red line. Performance of all the models was evaluated on a 1 June 2019–31 December 2020 dataset, which is completely independent of the training data.

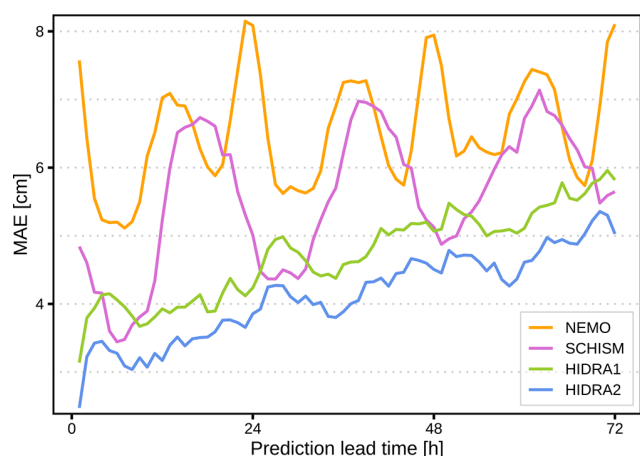


Figure 10. MAE score of the HIDRA, NEMO, and SCHISM models with regard to prediction lead time (between 1 and 72 h). Performance of all the models was evaluated on a 1 June 2019–31 December 2020 dataset, which is completely independent of the training data.

than NEMO but exhibits a similar periodicity in errors. Interestingly, while HIDRA2 consistently outperforms HIDRA1 for all lead times, the shapes of the MAE curves show a resemblance. While the 12 h period does not seem to be present in the MAE curves of these two models, their 24 h period is clearly present. Further research would, however, be required to substantiate and explain the observed MAE curve behavior.

4.2.3 Spectral analysis

To investigate the spectral properties of the modeled and observed SSH time series, we computed spectral densities of the HIDRA2, HIDRA1, NEMO and SCHISM predictions. Unless otherwise stated, all time series analyzed in this section were obtained by concatenating (in time) the first 24 h of each daily HIDRA2, HIDRA1 and NEMO 3 d forecast. Spectral densities (shown in Fig. 11) were then computed as absolute values of a 1D fast Fourier transform of the respective series over a fixed frequency domain of $(1\text{ h})^{-1}$ – $(72\text{ h})^{-1}$.

Figure 11 indicates that all methods adequately represent the tidal dynamics in Koper. The energy content around the two lowest basin eigenmodes is, however, more discriminatory: NEMO (Fig. 11a) clearly underestimates the spectral density both around the ground state seiche (at the 21.5 h period) and around the first excited state (10.9 h period). Similar behavior was noticed in our previous work with an independent configuration of NEMO (Žust et al., 2021). SCHISM, on the other hand, overestimates the energy in the ground state seiche band but reproduces the first excited state energy very well (Fig. 11b). HIDRA1 underestimates the energy of this part of the signal as well but nevertheless does a bit better by packing more energy density into these two bands. Predictions of HIDRA2 are clearly the closest to the

observations in the ground state seiche band but come a close second to SCHISM around the 10.9 h period.

It appears that HIDRA2 is capable of generating seiche-like behavior in its predictions. Spectral density, however, discards the temporal component of the signal, and adequate spectral density in the $(21.5\text{ h})^{-1}$ and $(10.9\text{ h})^{-1}$ frequency bands says little about whether Adriatic seiches are generated by HIDRA2 at the appropriate times, namely, during the storms. To inspect this aspect of HIDRA2 behavior, we now proceed to analyze the predictions during several historic storm tides.

4.2.4 Performance during historic storms

Historic Adriatic storm tide events are used to qualitatively compare the HIDRA2 performance with the state of the art. The storm tides in question occurred during November and December 2019 and were of historic proportions by any criterion. The Slovenian coast was flooded over 10 times in a single month, and sea levels in Venice were among the highest ever observed. Furthermore, the events in November 2019 turned out to be difficult to model due to the formation of a transient and very localized low pressure over the Gulf of Venice, which went unresolved in most models (Cavaleri et al., 2020). These events, along with those from December 2019, therefore represent a highly challenging benchmark for any atmospheric model and even more for any downstream SSH prediction method.

Figure 12 shows HIDRA2, HIDRA1, NEMO and SCHISM SSH forecasts for the Adriatic storm tide of November 2019. None of the models successfully predicted the first and highest sea level peak on 12 November 2019, but HIDRA2, NEMO and SCHISM all give a better forecast than HIDRA1, whose mean sea level does not even surpass the flooding threshold. As noted in Cavaleri et al. (2020), this peak was difficult to forecast due to the delicate timing between the peak of the winds and the peak of the full moon tide combined with the formation of an unresolved local pressure disturbance over the western coast of the northern Adriatic. The relative timing of these influences turned out to be a *sine qua non* for a successful prediction – neither the winds nor pressure were, in themselves, in any way extraordinary. It is further shown in Cavaleri et al. (2020) that this particular storm tide could have been up to 25 cm higher had this scenario evolved 12 h earlier, when tidal peaks were themselves higher.

The peak on 13 November is slightly better predicted by maximum members of both HIDRAs than by NEMO or SCHISM, with HIDRA2 exhibiting a somewhat lower forecast spread than HIDRA1. Apart from this peak, all the models captured the sea level variability quite well, which is in itself an implicit testament to the high skills of ECMWF atmospheric products.

The floods of December 2019 are another example of HIDRA2's superior performance over HIDRA1 and both

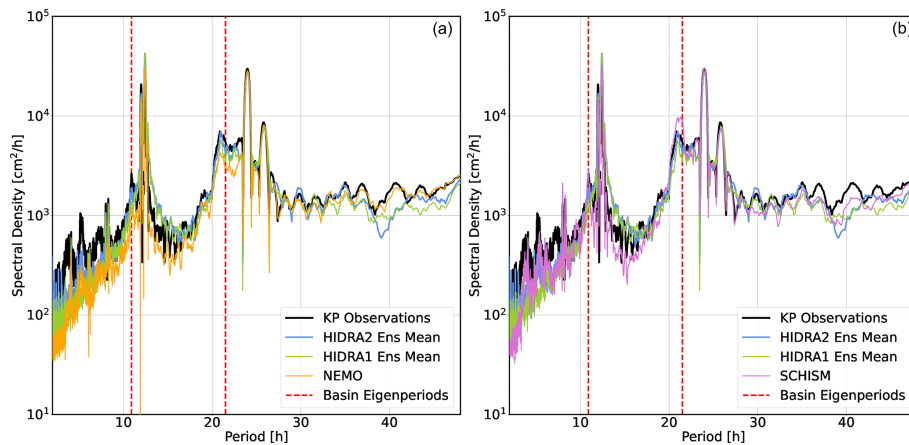


Figure 11. Spectral density of SSH time series from the Koper tide gauge, HIDRA2, and HIDRA1 compared with NEMO (a) and SCHISM (b) during independent cross-validation time windows between 1 June 2019 and 31 December 2020. Sharp peaks at (roughly) 12 and 24 h indicate the presence of tides, while the two dashed vertical red lines mark the periods of the two lowest Adriatic sea level eigenmodes. For clarity, all plotted spectral densities were filtered using a third-order Savitzky–Golay 24-point window filter.

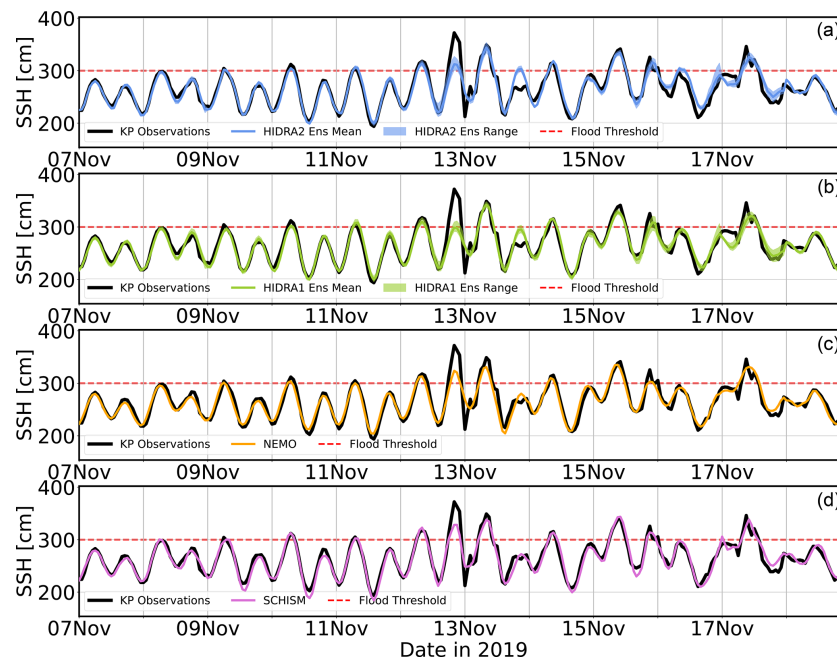


Figure 12. Comparison of the HIDRA2 ensemble (a), HIDRA1 ensemble (b), NEMO forecast run (c) and SCHISM reanalysis run (d) during the November 2019 flooding sequence in the northern Adriatic. Semi-transparent regions in (a) and (b) depict the minimum–maximum envelope of each HIDRA ensemble.

ocean models in Koper. SSH observations and predictions in Koper during this period are depicted in Fig. 13. Several conclusions about HIDRA2's behavior may be reached with regard to this particular flood. The HIDRA2 ensemble appears to be closest to the observations and exhibits a substantially lower forecasting spread than the HIDRA1 ensemble. Low forecasting spread is acceptable when in conjunction with a well-behaved ensemble mean. In this case, the HIDRA2 ensemble mean is in excellent agreement with the observations.

The same could be said for HIDRA1, albeit to a lesser degree. NEMO, however, completely misses the first two peaks between 15 and 17 December, slightly underestimates (like HIDRA1) the highest peak on 23 December, and overall underestimates the minimum–maximum range of the sea level variations, corresponding to poorly predicted ebb levels after 23 December. SCHISM predicts the first two peaks but underestimates the peaks after 23 December. The vertical sea level range is much better captured by both HIDRAs, espe-

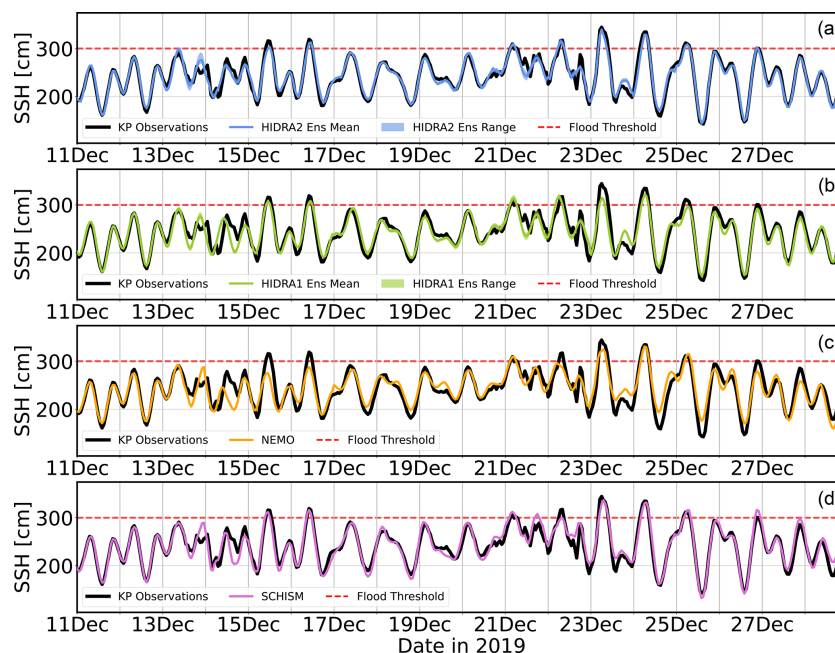


Figure 13. Same as Fig. 12 but for December 2019.

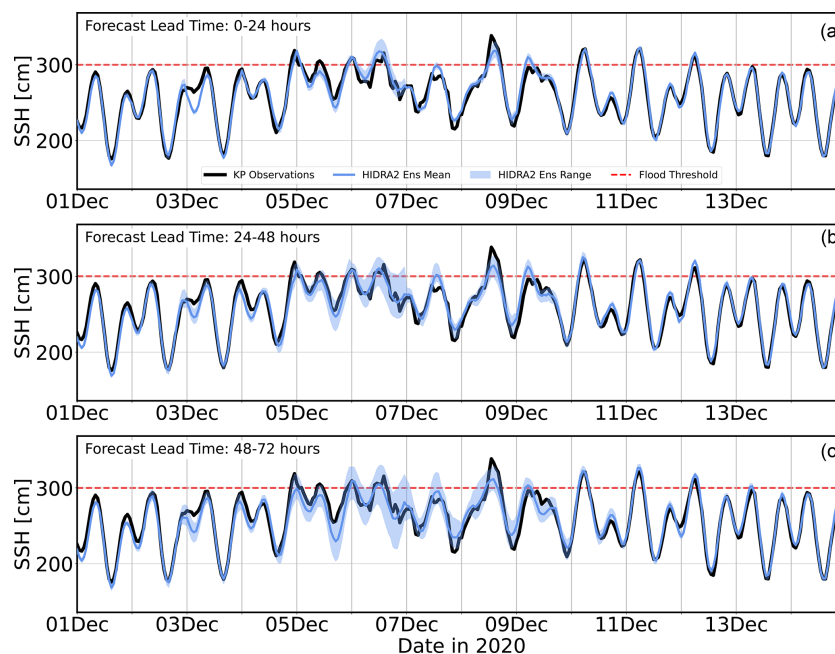


Figure 14. Comparison of the HIDRA2 ensemble spread on forecast days 1 (a), 2 (b) and 3 (c) during December 2020 in Koper. Semi-transparent regions in the plots depict the minimum–maximum range of the HIDRA2 ensemble.

cially by HIDRA2. This result is consistent with our demonstration that HIDRA2 exhibits the lowest error in both the high and low tails of sea level distributions (Fig. 11).

To inspect the behavior of the ensemble forecast spread, three time series were created from daily (72 h long) forecasts during the evaluation time window between 1 June 2019 and 31 December 2020. The first time series

was constructed by concatenating each first day (i.e., 1–24 h of forecast) from each of the daily forecasts, thus containing predictions with lead times of 1–24 h on each respective day in the evaluation time window. The second and third time series were constructed by concatenating 25–48 h (49–72 h) of forecast on each respective day in the evaluation time window. All three time series for the December 2020 floods are

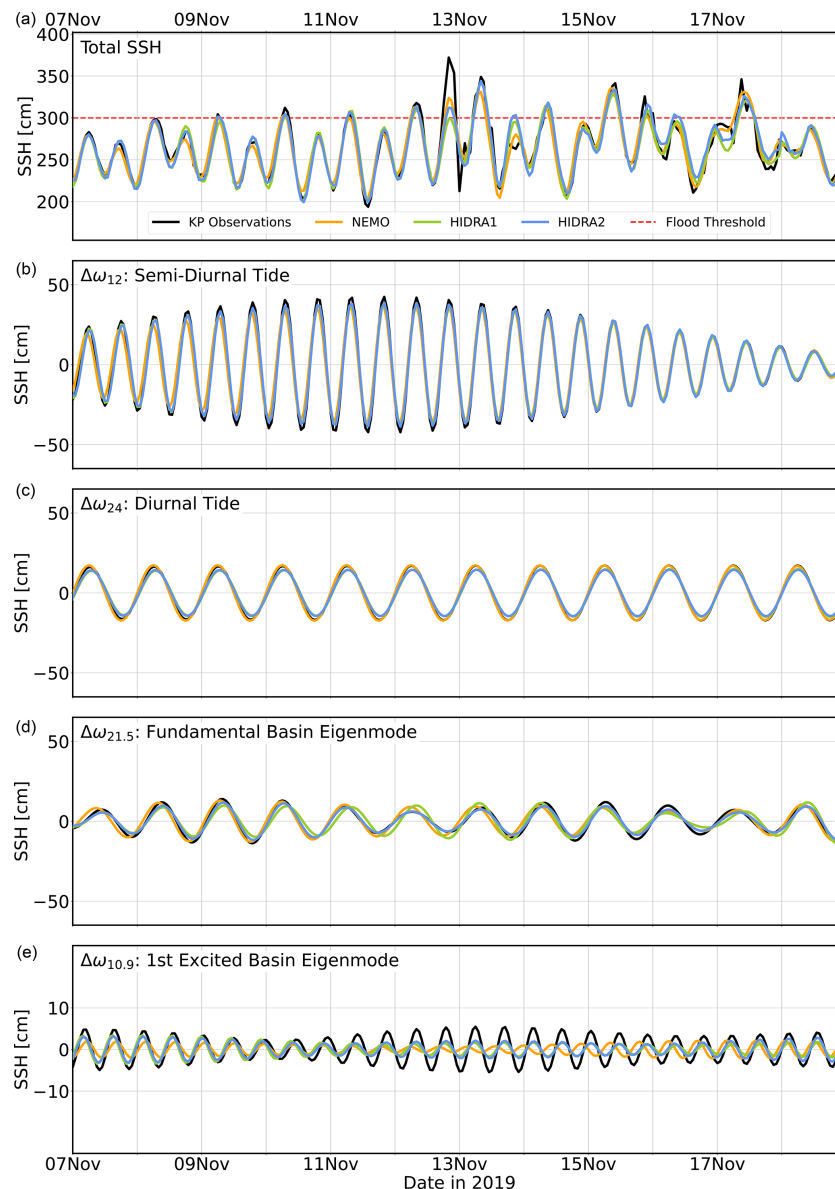


Figure 15. Comparison of total Koper SSH observations and forecasts (a) and their band-pass-filtered signals (b–e) over four bands, centered around four geophysically relevant periods (semi-diurnal and diurnal tides and the two lowest along-axis basin eigenmodes). The time window of the SSH signal spans from 7 November 2019 to 19 November 2019. Note the different vertical scale in (e).

shown in Fig. 14. As expected, from the growing ensemble spread in the atmospheric forcing, HIDRA2 spread is growing with forecast lead time as well. As we draw closer to a particular flooding event, the forecast spread drops, indicating an increased prediction certainty.

4.2.5 Spectral decomposition of forecasts during storms

To investigate the performance in geophysically relevant energy bands, we band-pass-filtered the observed and predicted SSH signals in energy bands, centered around four important

periods: semi-diurnal tide (12 h period), diurnal tide (24 h period), fundamental basin along-axis eigenmode (21.5 h period) and first excited along-axis eigenmode (10.9 h period).

Although incomplete, this SSH decomposition allows qualitative estimation of the excitation intensity of the basin eigenmodes during a particular storm and also helps to qualitatively assign forecasting errors to specific frequency bands. However, since the amplitudes of the filtered signals in Fig. 15 directly depend on the filter bandwidths, they should not be interpreted as direct contributions to the sea level due to respective geophysical phenomena (i.e., two tidal signals, two eigenmodes). They should rather be read strictly as an

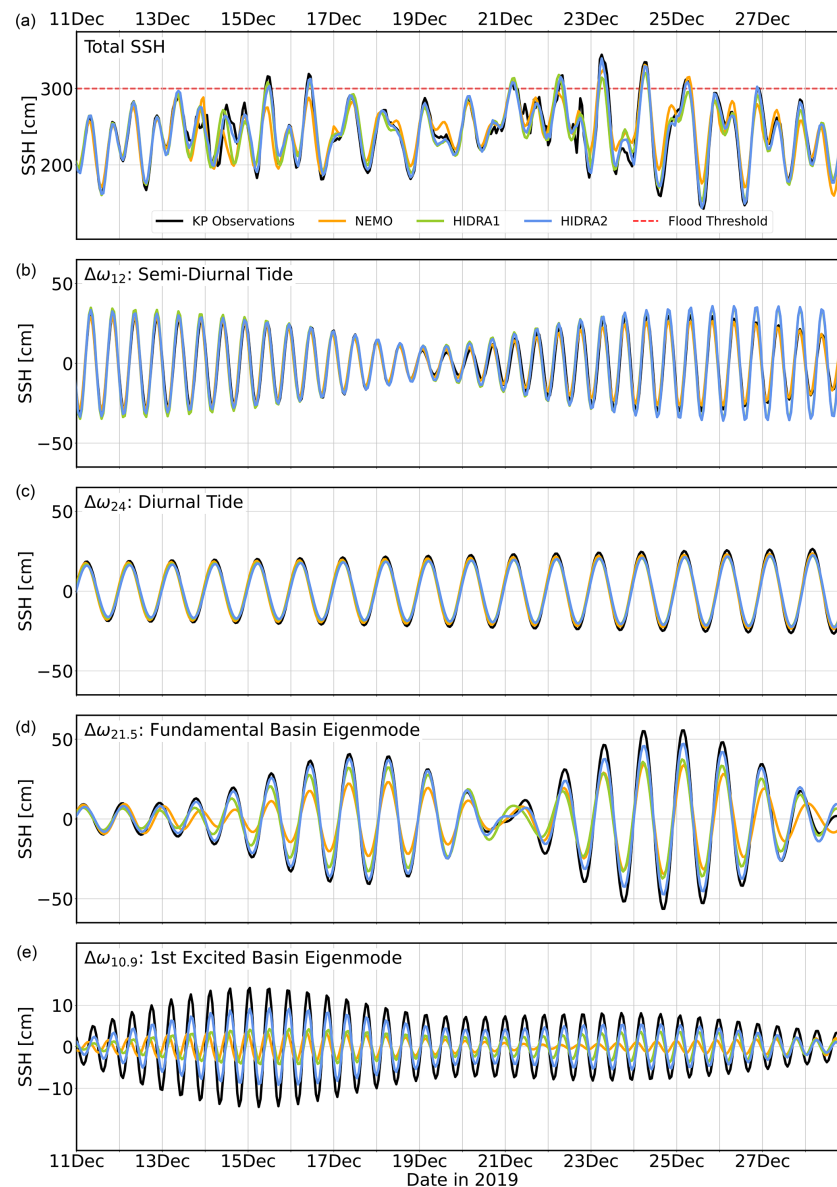


Figure 16. Same as Fig. 15 but for the December 2019 coastal floods.

additional insight into the model performance within a specific band with reference to filtered observations in the same band.

We applied a fifth-order Butterworth band-pass filter with a sampling rate of $(1\text{ h})^{-1}$. Low and high cutoff frequencies, which define the semi-diurnal filtering band $\Delta\omega_{12}$, were set to $\Delta\omega_{12} = [(12.5\text{ h})^{-1}, (11.5\text{ h})^{-1}]$. Similarly, diurnal cutoff frequencies were set to $\Delta\omega_{24} = [(24.5\text{ h})^{-1}, (23.5\text{ h})^{-1}]$. A fundamental seiche filtering band was estimated from Fig. 11 to be $\Delta\omega_{21.5} = [(20\text{ h})^{-1}, (24\text{ h})^{-1}]$, which is also consistent with the seiche window used in Vilibić (2006). Finally, the first excited eigenmode band is defined as $\Delta\omega_{10.9} = [(11.4\text{ h})^{-1}, (10.5\text{ h})^{-1}]$. An example of this decomposition for November 2019 is shown in Fig. 15. For brevity, we only

show results for the NEMO model in the main body of the paper.

Identical analysis and related figures for the SCHISM model are available in the Supplement to this paper. They illustrate that SCHISM exhibits very solid performance in the seiche energy bands.

All the models exhibit an underestimation of the amplitude but are otherwise in phase with the observations in the $\Delta\omega_{12}$ band. In $\Delta\omega_{24}$, NEMO seems to be performing very well, with HIDRA2 slightly underestimating the range of the signal in this band. In the band $\Delta\omega_{21.5}$ NEMO is again closest to filtered observations, while both HIDRA models overpredict the vertical range of the observed signal. Band $\Delta\omega_{10.9}$ is underpredicted in all the models but seems best (or rather

least poorly) resolved by HIDRAs, with NEMO additionally exhibiting a substantial phase shift in the signal.

In any case, since both tidal bands and the ground state seiche are reliably predicted by all the models, the reason for the forecasting errors must lie in the higher-frequency bands with periods below 10.9 h. This seems consistent with the occurrence of highly transient and localized low pressure over Venice mentioned in Cavaleri et al. (2020) and will be the subject of further research.

Similar remarks can be made regarding the December 2019 coastal flooding depicted in Fig. 16. This event marked a suboptimal performance of NEMO, which is systematically underestimating SSH peaks and the overall vertical range of the SSH variability during this time window (Fig. 16a). This caused NEMO to miss four floods out of eight. HIDRA models perform better, with HIDRA2 most reliably predicting all the flood peaks, most notably those on 15, 16 and 23 December 2019.

Figure 16b and c demonstrate that all the models are reliable in the diurnal tidal band $\Delta\omega_{24}$ but that HIDRA2 overestimates the signal in $\Delta\omega_{12}$. Since the overall performance of HIDRA2 is the best of all three models, it is unclear whether overshoots in $\Delta\omega_{12}$ could be interpreted as compensations for the underestimations in the nearby $\Delta\omega_{10.9}$ band. Figure 16d and e, however, indicate that some of the modeling errors stem from their underestimation of the basin seiches.

In the $\Delta\omega_{21.5}$ band, the HIDRA2 predictions most closely resemble the observations, followed by HIDRA1 and then NEMO (which is most severely underestimating this part of the signal). HIDRA2 is also the most reliable method in $\Delta\omega_{10.9}$ – but it nevertheless systematically underestimates the observations. HIDRA1 and NEMO performances are significantly worse, reaching one-half of the amplitude of HIDRA2 and one-third that of the observations. Poor performances of HIDRA1 and NEMO in the $\Delta\omega_{21.5}$ and $\Delta\omega_{10.9}$ bands are simply another reflection of the fact depicted in Fig. 11, namely, that both of these models struggle to generate an appropriate amount of energy in the bands around free oscillation eigenmodes.

5 Conclusions

This study presents a deep-learning-based sea level model, HIDRA2, suitable for operational sea level ensemble modeling due to its speed and accuracy. This work is a conceptual continuation of our previous attempt at sea level forecasting (Žust et al., 2021) and represents a substantial advancement over the first version (HIDRA1), setting a new state of the art in machine learning SSH forecasting. The new architecture is validated by extensive ablation studies. The performance is benchmarked against the current state-of-the-art Mediterranean forecasting setup of the NEMO ocean model (available as part of the Copernicus Marine Service) and against a multi-decadal reanalysis run of the SCHISM model

(Toomey et al., 2022) on an unstructured grid with very high coastal resolution. We demonstrate that HIDRA2 outperforms HIDRA1 as well as numerical ocean models across all sea level bins. We further show that HIDRA2 very accurately represents the energy contents in the bands around the relevant geophysical periods (diurnal and semi-diurnal tides and the lowest two free oscillation basin eigenmodes).

Performance is analyzed during several historic storms. Spectral decomposition of the total sea level signal into bands centered around tides and basin seiches is carried out to assign modeling errors to specific energy bands of the predicted sea levels. HIDRA2 consistently exhibits high skill in exciting the ground state Adriatic basin seiche at the appropriate time and with the appropriate phase and amplitude.

HIDRA2 is a good example of how the entanglement of deep learning and geophysics may lead to reliable and numerically cheap models that are able to mimic complex physical phenomena on the level of the best numerical physical models. Nevertheless, several extensions could be additionally explored. One possible extension is data ingestion from several tide gauges along the Adriatic coast and verification of whether the prediction accuracy at individual locations improves in such a multi-point prediction setup. Another extension is the inclusion of real-time in situ measurements such as synoptic observations and satellite scatterometer and wind measurements. It would be interesting to migrate HIDRA2 to other Mediterranean locations or other semi-enclosed basins like the Baltic Sea, the Red Sea or the Chesapeake Bay to investigate its generalization properties. These will be the objects of our future research.

Code and data availability. Implementation of HIDRA2 and the code to train and evaluate the model are available in the GitHub repository: <https://github.com/rusmarko/HIDRA2> (last access: 9 November 2022). We also include HIDRA2 weights pretrained on 2006–2018 and predictions for all 50 ensembles in June 2019–December 2020. The persistent version of our GitHub repository containing code is available at <https://doi.org/10.5281/zenodo.7307365> (Rus et al., 2022a). We publish our training and evaluation datasets at <https://doi.org/10.5281/zenodo.7304086> (Rus et al., 2022c). Sea level datasets employed in this paper are available at <https://doi.org/10.5281/zenodo.7277108> (Rus et al., 2022b).

Supplement. The supplement related to this article is available online at: <https://doi.org/10.5194/gmd-16-271-2023-supplement>.

Author contributions. MR was the main designer of HIDRA2 and reimplemented HIDRA1 in PyTorch. MK led the machine learning part of the research and contributed to the design of HIDRA2. ML provided the geophysical background relevant for the HIDRA2 design and led the geophysical part of the research. AF and ML prepared the atmospheric and sea level training and evaluation datasets.

MR, MK and ML analyzed the results and wrote the paper. All the authors contributed to the final version of the manuscript.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. The authors would like to thank Tim Toomey and Alejandro Orfila for providing SCHISM sea level reanalysis time series for the Koper location, which were used for benchmarking HIDRA2 in this work. We further thank the reviewers for taking the time to review the manuscript and for their constructive remarks which led to an improved paper. Matjaž Ličer was financially supported by the Slovenian Research Agency (research core funding no. P1-0237). Matej Kristan was financially supported by the Slovenian Research Agency (research project no. J2-2506).

Financial support. This research has been supported by the Slovenian Research Agency (research core funding no. P1-0237 and research project no. J2-2506).

Review statement. This paper was edited by Rohitash Chandra and reviewed by three anonymous referees.

References

- Adams, D.: The Hitchhiker's Guide to the Galaxy, 42nd edn., Pan-Macmillan, ISBN 978-1-5290-4613-7, 1979.
- Arias, P. A., Bellouin, N., Coppola, E., Jones, R. G., Krinner, G., Marotzke, J., Naik, V., Palmer, M. D., Plattner, G.-K., Rogelj, J., Rojas, M., Sillmann, J., Storelvmo, T., Thorne, P. W., Trewin, B., Achuta Rao, K., Adhikary, B., Allan, R. P., Armour, K., Bala, G., Barimalala, R., Berger, S., Canadell, J. G., Cassou, C., Cherchi, A., Collins, W., Collins, W. D., Connors, S. L., Corti, S., Cruz, F., Dentener, F. J., Dereczynski, C., Di Luca, A., Diongue Niang, A., Doblas-Reyes, F. J., Dosio, A., Duville, H., Engelbrecht, F., Eyring, V., Fischer, E., Forster, P., Fox-Kemper, B., Fuglestad, J. S., Fyfe, J. C., Gillett, N. P., Goldfarb, L., Gorodetskaya, I., Gutierrez, J. M., Hamdi, R., Hawkins, E., Hewitt, H. T., Hope, P., Islam, A. S., Jones, C., Kaufman, D. S., Kopp, R. E., Kosaka, Y., Kossin, J., Krakovska, S., Lee, J.-Y., Li, J., Mauritsen, T., Maycock, T. K., Meinshausen, M., Min, S.-K., Monteiro, P. M. S., Ngo-Duc, T., Otto, F., Pinto, I., Pirani, A., Raghavan, K., Ranasinghe, R., Ruane, A. C., Ruiz, L., Sallée, J.-B., Samset, B. H., Sathyendranath, S., Seneviratne, S. I., Sörensson, A. A., Szopa, S., Takayabu, I., Treguier, A.-M., van den Hurk, B., Vautard, R., von Schuckmann, K., Zaehle, S., Zhang, X., and Zickfeld, K.: Technical Summary, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., book section 1, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_TS.pdf (last access: 23 December 2022), 2021.
- Bajo, M., Medugorac, I., Umgiesser, G., and Orlić, M.: Storm surge and seiche modelling in the Adriatic Sea and the impact of data assimilation, *Q. J. Roy. Meteor. Soc.*, 145, 2070–2084, <https://doi.org/10.1002/qj.3544>, 2019.
- Bernier, N. B. and Thompson, K. R.: Deterministic and ensemble storm surge prediction for Atlantic Canada with lead times of hours to ten days, *Ocean Model.*, 86, 114–127, <https://doi.org/10.1016/j.ocemod.2014.12.002>, 2015.
- Braakmann-Folgmann, A., Roscher, R., Wenzel, S., Uebbing, B., and Kusche, J.: Sea level anomaly prediction using recurrent neural networks, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.1710.07099>, 19 October 2017.
- Cavaleri, L., Bajo, M., Barbariol, F., Bastianini, M., Benetazzo, A., Bertotti, L., Chiggiato, J., Ferrarin, C., and Umgiesser, G.: The 2019 Flooding of Venice and Its Implications for Future Predictions, *Oceanography*, 33, 42–49, <https://doi.org/10.5670/oceanog.2020.105>, 2020.
- Cerovecki, I., Orlic, M., and Hendershott, M. C.: Adriatic seiche decay and energy loss to the Mediterranean, *Deep-Sea Res. Pt. I*, 44, 2007–2029, [https://doi.org/10.1016/s0967-0637\(97\)00056-3](https://doi.org/10.1016/s0967-0637(97)00056-3), n/a, 1997.
- Clementi, E., Aydogdu, A., Goglio, A. C., Pistoia, J., Escudier, R., Drudi, M., Grandi, A., Mariani, A., Lyubartsev, V., Lecci, R., Cretí, S., Coppini, G., Masina, S., and Pinardi, N.: Mediterranean Sea Analysis and Forecast (CMEMS MED-Currents, EAS6 system) (Version 1), Copernicus Monitoring Environment Marine Service (CMEMS) [data set], https://doi.org/10.25423/cmcc/medsea_analysis_forecast_phy_006_013_eas4, 2021.
- Codiga, D.: Unified Tidal Analysis and Prediction Using the UTide Matlab Functions, Tech. Rep., Graduate School of Oceanography, University of Rhode Island, Narragansett, RI, USA, GitHub [code], <https://github.com/wesleybowman/UTide> (last access: 14 November 2022), 2011.
- Ferrarin, C., Valentini, A., Vodopivec, M., Klaric, D., Massaro, G., Bajo, M., De Pascalis, F., Fadini, A., Ghezzi, M., Mene-gon, S., Bressan, L., Unguendoli, S., Fettich, A., Jerman, J., Ličer, M., Fustar, L., Papa, A., and Carraro, E.: Integrated sea storm management strategy: the 29 October 2018 event in the Adriatic Sea, *Nat. Hazards Earth Syst. Sci.*, 20, 73–93, <https://doi.org/10.5194/nhess-20-73-2020>, 2020.
- Ferrarin, C., Lionello, P., Orlić, M., Raicich, F., and Salvadori, G.: Venice as a paradigm of coastal flooding under multiple compound drivers, *Scientific Reports*, 12, 5754, <https://doi.org/10.1038/s41598-022-09652-5>, 2022.
- Gregory, J., S.M., G., Hughes, C., Lowe, J. A., Church, J. A., Fukimori, I., Gomez, N., Kopp, R. E., Landerer, F., Le Cozannet, G., Ponte, R. M., Stammer, D., Tamisiea, M. E., and van de Wal, R. S. W.: Concepts and Terminology for Sea Level: Mean, Vari-

- ability and Change, Both Local and Global, *Surv. Geophys.*, 40, 1251–1289, <https://doi.org/10.1007/s10712-019-09525-z>, 2019.
- He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, 27–30 June 2016, 770–778, <https://doi.org/10.1109/CVPR.2016.90>, 2016.
- Hieronymus, M., Hieronymus, J., and Hieronymus, F.: On the application of machine learning techniques to regression problems in sea level studies, *J. Atmos. Ocean. Tech.*, 36, 1889–1902, 2019.
- Hochreiter, S. and Schmidhuber, J.: Long short-term memory, *Neural Comput.*, 9, 1735–1780, 1997.
- Imani, M., Kao, H.-C., Lan, W.-H., and Kuo, C.-Y.: Daily sea level prediction at Chiayi coast, Taiwan using extreme learning machine and relevance vector machine, *Global Planet. Change*, 161, 211–221, 2018.
- Ioffe, S. and Szegedy, C.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in: Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 7–9 July 2015, edited by: Bach, F. and Blei, D., PMLR, 37, 448–456, <http://proceedings.mlr.press/v37/loff15.pdf> (last access: 14 November 2022), 2015.
- Ishida, K., Tsujimoto, G., Ercan, A., Tu, T., Kiyama, M., and Amagasaki, M.: Hourly-scale coastal sea level modeling in a changing climate using long short-term memory neural network, *Sci. Total Environ.*, 720, 137613, <https://doi.org/10.1016/j.scitotenv.2020.137613>, 2020.
- Karimi, S., Kisi, O., Shiri, J., and Makarynsky, O.: Neuro-fuzzy and neural network techniques for forecasting sea level in Darwin Harbor, Australia, *Comput. Geosci.*, 52, 50–59, 2013.
- Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S.: Self-normalizing neural networks, *Adv. Neur. In.*, 30, 971–980, 2017.
- Leutbecher, M. and Palmer, T.: Ensemble forecasting, *Tech. Rep.*, ECMWF, <https://doi.org/10.21957/c0hq4yg78>, 2007.
- Loshchilov, I. and Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.1608.03983>, 13 August 2016.
- Loshchilov, I. and Hutter, F.: Decoupled weight decay regularization, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.1711.05101>, 14 November 2017.
- Madec, G.: NEMO ocean engine, Note du Pôle de modélisation, Institut Pierre-Simon Laplace (IPSL), France, No. 27, ISSN No 1288-1619, https://www.nemo-ocean.eu/wp-content/uploads/NEMO_book.pdf (last access: 14 November 2022), 2016.
- Medvedev, I. P., Vilibić, I., and Rabinovich, A. B.: Tidal resonance in the Adriatic Sea: Observational evidence, *J. Geophys. Res.-Oceans*, 125, e2020JC016168, <https://doi.org/10.1029/2020JC016168>, 2020.
- Mel, R. and Lionello, P.: Storm Surge Ensemble Prediction for the City of Venice, *Weather Forecast.*, 29, 1044–1057, <https://doi.org/10.1175/WAF-D-13-00117.1>, 2014.
- Pashova, L. and Popova, S.: Daily sea level forecast at tide gauge Burgas, Bulgaria using artificial neural networks, *J. Sea Res.*, 66, 154–161, 2011.
- Pérez Gómez, B., Vilibić, I., Šepić, J., Međugorac, I., Ličer, M., Testut, L., Fraboul, C., Marcos, M., Abdellaoui, H., Álvarez Fanjul, E., Barbalić, D., Casas, B., Castaño-Tierno, A., Čupić, S., Drago, A., Fraile, M. A., Galliano, D. A., Gauci, A., Gloginja, B., Martín Guijarro, V., Jeromel, M., Larrad Revuelto, M., Lazar, A., Keskin, I. H., Medvedev, I., Menassri, A., Meslem, M. A., Mihanović, H., Morucci, S., Niculescu, D., Quijano de Benito, J. M., Pascual, J., Palazov, A., Picone, M., Raichich, F., Said, M., Salat, J., Sezen, E., Simav, M., Sylaios, G., Tel, E., Tintoré, J., Zaimi, K., and Zodiatis, G.: Coastal sea level monitoring in the Mediterranean and Black seas, *Ocean Sci.*, 18, 997–1053, <https://doi.org/10.5194/os-18-997-2022>, 2022.
- Rus, M., Fettich, A., Kristan, M., and Ličer, M.: Code for HIDRA2: Deep-Learning Ensemble Storm Surge Forecasting in the Presence of Seiches – the Case of Northern Adriatic, Zenodo [code], <https://doi.org/10.5281/zenodo.7307365>, 2022a.
- Rus, M., Fettich, A., Kristan, M., and Ličer, M.: Sea Level Datasets for HIDRA2 Training and Evaluation, Zenodo [data set], <https://doi.org/10.5281/zenodo.7277108>, 2022b.
- Rus, M., Fettich, A., Kristan, M., and Ličer, M.: Training and Test Datasets for HIDRA2, Zenodo [data set], <https://doi.org/10.5281/zenodo.7304086>, 2022c.
- Sapankevych, N. I. and Sankar, R.: Time Series Prediction Using Support Vector Machines: A Survey, *IEEE Comput. Intell. M.*, 4, 24–38, <https://doi.org/10.1109/MCI.2009.932254>, 2009.
- Sonnewald, M., Lguensat, R., Jones, D. C., Dueben, P. D., Brajard, J., and Balaji, V.: Bridging observations, theory and numerical simulation of the ocean using machine learning, *Environ. Res. Lett.*, 16, 073008, <https://doi.org/10.1088/1748-9326/ac0eb0>, 2021.
- Taherkhani, M., Vitousek, S., Barnard, P., Frazer, N., Anderson, T. R., and Fletcher, C. H.: Sea-level rise exponentially increases coastal flood frequency, *Scientific Reports*, 10, 6466, <https://doi.org/10.1038/s41598-020-62188-4>, 2020.
- Toomey, T., Amores, A., Marcos, M., and Orfila, A.: Coastal sea levels and wind-waves in the Mediterranean Sea since 1950 from a high-resolution ocean reanalysis, *Frontiers in Marine Science*, 9, <https://doi.org/10.3389/fmars.2022.991504>, 2022.
- Vilibić, I.: The role of the fundamental seiche in the Adriatic coastal floods, *Cont. Shelf Res.*, 26, 206–216, <https://doi.org/10.1016/j.csr.2005.11.001>, 2006.
- Zhang, Y. J., Ye, F., Stanev, E. V., and Grashorn, S.: Seamless cross-scale modeling with SCHISM, *Ocean Model.*, 102, 64–81, <https://doi.org/10.1016/j.ocemod.2016.05.002>, 2016.
- Žust, L., Fettich, A., Kristan, M., and Ličer, M.: HIDRA 1.0: deep-learning-based ensemble sea level forecasting in the northern Adriatic, *Geosci. Model Dev.*, 14, 2057–2074, <https://doi.org/10.5194/gmd-14-2057-2021>, 2021.