Hindcast of Significant Wave Heights in Sheltered Basins Using Machine Learning and the Copernicus Database

Bujak, Damjan; Carević, Dalibor; Bogovac, Tonko; Kulić, Tin

Source / Izvornik: **Naše more : znanstveni časopis za more i pomorstvo, 2023, 70, 103 - 114**

Journal article, Published version
Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

https://doi.org/10.17818/NM/2023/2.5

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:237:751620

Hindcast of Significant Wave Heights in Sheltered Basins Using Machine Learning and the Copernicus Database

Kratkoročna prognoza značajnih valnih visina u zaštićenim akvatorijima koristeći se strojnim učenjem i Copernicus bazom podataka

Damjan Bujak*
University of Zagreb
Faculty of Civil Engineering
E-mail: damjan.bujak@grad.unizg.hr

Dalibor Carević
University of Zagreb
Faculty of Civil Engineering
E-mail: dalibor.carevic@grad.unizg.hr

Tonko Bogovac
University of Zagreb
Faculty of Civil Engineering
E-mail: tonko.bogovac@grad.unizg.hr

Tin Kulić
University of Zagreb
Faculty of Civil Engineering
E-mail: tin.kulic@grad.unizg.hr

DOI 10.17818/NM/2023/2.5
UDK 551.466:004.85
004.032.2
Original scientific paper / *Izvorni znanstveni rad*
Paper received / *Rukopis primljen:* 12. 7. 2022.
Paper accepted / *Rukopis prihvaćen:* 20. 4. 2023.

Abstract

Long-term time series of wave parameters play a critical role in coastal structure design and maritime activities. At sites with limited buoy measurements, methods are used to extend the available time series data. To date, wave hindcasting research using machine learning methods has mainly focused on filling in missing buoy measurements or finding a mapping function between two nearshore buoy locations. This work aims to implement machine learning methods for hindcasting wave parameters using only publicly available Copernicus data. Ensemble regression and artificial neural networks were used as machine learning methods and the optimal hyperparameters were determined by the Bayesian optimization algorithm. As inputs, data from the MEDSEA reanalysis wave model were used for the wave parameters and data from the ERA5 atmospheric reanalysis model were used for the wind parameters. The results of this study show that the normalized *RMSE* of the test data improved by 29% for Rijeka and 12% for Split compared to the original MEDSEA wave hindcast at buoy locations. The proposed method was extremely efficient in removing bias in the original MEDSEA hindcasts (e.g., *NBIAS* = -0.35 for Rijeka) to negligible values for both Split and Rijeka (*NBIAS* < 0.03).

Sažetak

Dugoročne vremenske serije parametara vala igraju značajnu ulogu u projektiranju pomorskih građevina i pomorskim aktivnostima. Na mjestima s ograničenim direktnim mjerenjima plutača metode se koriste kako bi se proširio vremenski niz raspoloživih podataka. Do danas, istraživanja o uspostavi kratkoročnih prognoza korištenjem metodama strojnog učenja uglavnom se usredotočilo na nadomještanje nedostajućih mjerenih podataka plutače ili pronalazak funkcijske veze između dvaju mjesta plutača u blizini obale. Ovaj rad ima za cilj koristiti metode strojnoga učenja radi provedbe kratkoročnih prognoza koristeći se samo javnodostupnim podacima Copernicus. Ansambl regresija i umjetne neuralne mreže koriste se kao metode strojnoga učenja, a optimalni hiperparametri određeni su Bayesianovim algoritmom optimizacije. Podaci valova iz MEDSEA modela i podaci vjetra iz ERA5 modela atmosfere su korišteni kao ulazni podaci. Rezultati ove studije pokazuju da je za testni set RMSE se smanjio za 29% za Rijeku i 12% za Split uspoređujući s izvornom MESEA kratkoročnom prognozom valova na lokacijama plutača. Predložena metoda bila je izuzetno djelotvorna pri uklanjanju pristranosti u izvornoj MEDSEA kratkoročnoj prognozi (npr. = -0,35 za Rijeku) do zanemarivih vrijednosti i za Split i za Rijeku (NBIAS < 0.03).

KEY WORDS
machine learning
significant wave height
ANN
ensemble regression
CMEMS

KLJUČNE RIJEČI
strojno učenje
značajna visina vala
ANN
regresija sklopa
CMMES

1. INTRODUCTION / *Uvod*

Knowledge of a long-term time series of wave climate (e.g. significant wave height, peak wave period, etc.) at a location is essential for planning, operation, and maintenance of maritime activities [1], flood protection engineering design [2], and coastal vulnerability assessment [3]. A long wave height time series is widely recognized as key to reliable long-term significant wave height forecasting or hindcasting, (e.g., to define the return level period of significant wave height) [4]. This long-term wave height time series can be constructed using wave hindcast methods when data is missing or not sufficient [5]. The high return period of significant wave heights (50-year, 100-year return period, etc.) is required as input for the planning phase of coastal structures, long-term morphodynamics process studies [6], or as a boundary value for numerical nearshore wave climate models [7]. The capability and availability of a reliable long-term wave database are of paramount importance to the ocean and coastal engineers. This is especially true for coastal seas, for which there are much less data compared to deeper seas.

* Corresponding author

Wave buoys provide long-term time series of wave parameters at a variety of locations if national or international climate monitoring networks maintain them continuously. If this is the case, wave records could be available for 15-30 years [8]. In the Mediterranean Sea, for example, the Hellenic Centre for Marine Research, and Institute of Oceanography started monitoring the wave field with their buoy ATHOS in May of 2000, and the Spanish Harbor Authority with their buoy 6100197 in April of 1993 [9]. In addition, there are 39 years of continuous wave measurements at the Aqua Alta oceanographic tower in the Adriatic Sea [10]. More commonly, however, wave buoys are only maintained for only a few years during specific campaigns before they are usually recovered. This does not provide long enough time series for long-term forecasting, so methods to extend the time series should be applied [11].

One way to increase the available data set at sites where measurements have been made for only a limited duration is to use historical wind data as input to wave hindcasting to reconstruct a long-term wave time series. There is a wide range of solutions for this, ranging from simple empirical models [11; 12] to complex wave-generating numerical models [13; 14]. Numerical models are limited by available computational power, detailed bathymetry data at locations sheltered by islands, their complexity, and difficult-to-determine coefficient (e.g. white-capping parameters, bed frictional dissipation, depth-limited wave breaking, etc.) [15], while the main advantage compared to the simpler empirical models is that they perform physically based calculations. Hindcasting of wave heights at a local level from global/regional reanalysis models has also been performed using locally based wave numerical models (SWAN) using the reanalysis data as input [7; 16].

Climate reanalysis products can provide data for historical wind and wave data for a specific location or an entire region. The website Advancing Reanalysis [17] provides a visual comparison of the various reanalysis products [18]. Reanalysis is a scientific method of creating a complete record of changes in weather and climate over time. It usually spans decades or more and covers the entire Earth or focuses on a specific region. The information obtained from reanalysis is widely used for monitoring, comparison, determining the causes of climate variability, and supplementing climate predictions. The Copernicus database provides several reanalysis models that include wave data, such as ERA5 global wave climate [19], WAVERYS - Global Ocean Waves Reanalysis [20], and MEDSEA - Mediterranean Sea Waves Reanalysis [21]. WAVERYS and ERA5 both have wave height grided at low resolution (0.2° and 0.25°, respectively) when compared to MEDSEA (1/24°). Furthermore, the WAVERYS model reanalysis has a temporal resolution of 3 h, compared to the 1 h temporal resolution of MEDSEA and ERA5. The disadvantage of the MEDSEA regional model compared to the other two is its relatively short time range 1993-2020 (last accessed on 27.04.2022.), while the ERA5 model has data back to 1979. A longer wave and wind history could contribute to a longer wave time series when hindcasting. In addition to the wave data provided, ERA5 also offers wind reanalysis data at a spatial resolution of 0.25°. Although the MEDSEA reanalysis model is the most detailed Copernicus numerical model reanalysis in the Mediterranean Sea (both in spatial and temporal terms), Korres et al. [21] still observed low accuracy when validating the reanalysis data to buoy measurements in well-sheltered areas. They note that the worst model performance is observed for the Adriatic Sea. Korres et al. [21] argue that the Adriatic Sea is shallow, enclosed, and bounded by complex topography, and therefore not adequately represented by the spatial resolution of the forcing wind and possibly by the spatial resolution of the wave model.

There is an alternative to performing hindcasting using complex numerical reanalysis models, such as hindcasting using machine learning models, from simple models such as stepwise linear models to complex artificial neural networks (ANN), and more. Machine learning models are capable of mapping complex non-linear functions between inputs and outputs when sufficient training data is available [22]. These methods are used in many applications in the ocean and coastal engineering in many applications, such as wave forecasting with several hours of lead time [1; 15; 23; 24], wave runup [25], beach sediment transport [26; 27], beach nourishment requirements [28], etc. As for wave hindcasting, it has been mainly performed using ANN based on wind reanalysis data at a regional level to downscale it to a local level [8; 29]. Machine learning models have been used when data were missing in the measured wave time series to fill in missing wave heights [30] or to find a mapping function between wave data at a nearshore site, by using data from one or more nearby offshore sites [31]. To train and test a machine learning model of this kind, data from a short-term wave buoy campaign can be used. The insights gained from machine learning techniques can additionally be used to improve the hindcast predictions given by a reanalysis wave model to better fit the available campaign wave buoy measurements.

This study aims to present a modeling chain that uses machine learning models with state-of-the-art regional reanalysis wave data (MEDSEA) and global reanalysis wind data (ERA5) as input for long-term reconstruction of significant wave heights at three different locations in the Adriatic Sea, which has proven to be the most challenging region for MEDSEA. We propose a method to rapidly improve MEDSEA wave prediction at sheltered locations using machine learning tools. Consequently, a validated machine learning model can extend the wave time series beyond the duration of the buoy measurement at the site, which is initially available to the user. The paper evaluates two different machine learning techniques: artificial neural networks (ANN) and ensemble learning with regression trees. Validity is demonstrated by comparing the newly created hindcast using a machine learning model with the MEDSEA hindcast time series as a benchmark. This research explores the possibility of using a rapid methodology to extend data from wave monitoring campaigns to a time that has not been measured in the past. A reliably extended long-term wave time series is needed for predicting the return period from 1 to 100 years and consequently for planning coastal structures. In addition, the wave time series can be extended by periodically updating the input data set as new MEDSEA and ERA5 data become available over time.

This article is organized as follows: the methodology is outlined in section 2, and the paper will then go on to compare the MEDSEA modeled data with measured data in section 3.1 and analyze the influence of the machine learning correction method in section 3.2. Section 4 will present the paper's conclusions.

2. METHODOLOGY / *Metodologija*

2.1. CMEMS numerical wave model data set and ECMWF Reanalysis v5 (ERA5) numerical wind model data set / *Podaci valova CMEMS numeričkog modela i podaci vjetra (ERA5) numeričkog modela*

2.1.1. Wave data / *Podaci o valu*

The Copernicus Marine Environment Monitoring Service (CMEMS) provides a 27-year wave reanalysis product, MEDSEA, covering the period from January 1993 to December 2019 for the Mediterranean Sea [21]. This wave reanalysis is based on the advanced third-generation wave model WAM Cycle 4.6.2 [13; 32]. It explicitly solves the wave transport equations without assuming the wave spectrum shape. Included source terms are wind input, white capping dissipation, nonlinear transfer, and bottom friction. The wind and white-capping dissipation terms are based on Janssen's quasilinear theory of wind-wave generation [33; 34], while the bottom friction term is based on the empirical JONSWAP model [35]. The numerical model discretizes the wave spectra using 32 frequencies covering a logarithmically scaled frequency band from 0.04177 Hz to 0.8018 Hz (with wave periods from about 1 s to 24 s) and 24 uniformly distributed directional bins (bin size of 15 deg). Winds from the ERA5 reanalysis 10 m above the sea surface (Copernicus Climate Service - ECMWF) are forcing the numerical wave model. The bathymetric map was created using the GEBCO bathymetric dataset [36]. In addition, the reanalysis includes an assimilation scheme that uses the significant wave heights determined from altimeters and adjusts the wave spectrum at each grid point accordingly (originally developed by [37]).

Korres et al. [21] show typical differences between the MEDSEA reanalysis model and the in-situ and satellite observations (*RMSE*) of 0.23 ± 0.012 m and 0.24 ± 0.01 m respectively, and *BIAS* of -0.06 ± 0.022 m (7% ± 3% relative to the observed mean) and -0.05 ± 0.011 m (4% ± 1%) for the Mediterranean Sea as a whole. *BIAS* is predominately negative, indicating widespread underestimation of the measured wave heights by the reanalysis. In the Adriatic Sea, the model accuracy for buoy 61217 deteriorates further (*RMSE*) to 0.27 m and *BIAS* to -0.14, as stated by the model authors (Table 1).

The MEDSEA reanalysis model provides 2D hourly instantaneous fields (Table 2). It generates data every hour with a horizontal resolution of 1/24°. This dataset will be used as inputs to the machine learning model to hindcast wave heights that are more accurate to measured wave data at the Croatian coast (described in section 2.2.).

2.1.2. Wind data / *Podaci o vjetru*

ERA5 is the fifth generation of ECMWF's global climate and weather reanalysis for recent decades (starting in 1979). The reanalysis integrates model data with observations from around the world to produce a complete and consistent global dataset. This reanalysis also uses data assimilation, which optimally combines the forecast with newly available observations every 12 hours to produce an optimal estimate of atmospheric conditions. Because the reanalysis is not forced to produce timely forecasts, there is more time to collect new data and incorporate historical data to improve the quality of the reanalysis product. ERA5 provides hourly estimates for a variety of atmospheric, ocean wave, and land surface parameters. The wind data used in this work were resampled to a regular lat-lon grid with a resolution of 0.25°. Recent studies have compared the ERA5 reanalysis wind data with measured wind station data in the Adriatic and Mediterranean Seas. They showed moderate accuracy of ERA5 when compared to more detailed regional wind reanalysis, and significant underprediction (2 m/s on average) for high wind velocities (larger than 10 m/s) [38; 39]. Nevertheless, the ERA5 is a state-of-the-art atmosphere reanalysis model, and is publicly available, which is important from a data availability standpoint.

Lag components of significant wave height from CMEMS and wind magnitude from ECMWF were added to the predictor set to account for the historical components of these variables. Thus, at a given time in the reconstruction of the measured significant wave height, the model has insight into the current wave height hindcast from MEDSEA and 10 previous predictions of wave height (MEDSEA) and wind magnitude (ERA5). Similarly, some researchers have added this to help predict wave height [8; 23]. In this way, wind duration is taken into account, which is expected to affect the wave height reconstruction accuracy.

Table 1 Stations in the western Adriatic Sea (Italy) for which model validation and performance metrics were conducted in [21]; location of the wave buoys is shown in Figure 1 (red dots); 'O' and 'P' indicate observations and predictions on the station, respectively

Tablica 1. Postaje na zapadnome Jadranskome moru (Italija) za koje je izvršena validacija modela i metrika izvedbe [21]; lokacija plutača prikazana je na Slici 1 (crvene točke); 'O' i 'P' označavaju mjerenja i predviđanja na lokacijama postaja

Station	Years active	MEAN(O) (m)	MEAN(P) (m)	STD(O) (m)	STD(P) (m)	RMSE (m)	NRMSE
61217	1993-2011	0.64	0.5	0.55	0.44	0.27	0.42
61218	1999-2014	0.75	0.62	0.6	0.52	0.26	0.35
61220	2002-2014	0.56	0.44	0.5	0.43	0.22	0.39
ADN_DWRG1	2016-active	0.43	0.35	0.37	0.35	0.18	0.42
ADN_DWRG2	2016-active	0.33	0.26	0.27	0.25	0.14	0.42

Station	Years active	SI	BIAS (m)	NBIAS	CORR
61217	1993-2011	0.36	-0.14	-0.22	0.92
61218	1999-2014	0.30	-0.13	-0.17	0.93
61220	2002-2014	0.34	-0.11	-0.20	0.93
ADN_DWRG1	2016-active	0.37	-0.08	-0.19	0.90
ADN_DWRG2	2016-active	0.39	-0.06	-0.18	0.89

Table 2 Variable names and description from the CMEMS numerical wave model (variables 1-17) and ECMWF Reanalysis v5 (ERA5) wind model (variables 18-19)

Tablica 2. Nazivi varijabli i opis iz CMEMS numeričkoga modela valova (varijabla 1-17) i ECMWF reanaliza v5 (ERA5) modela vjetra (varijable 18-19)

Number	Variable Name	Description
1	VHM0	Spectral significant wave height (Hm0)
2	VHM0_SW1	Spectral significant primary swell wave height
3	VHM0_SW2	Spectral significant secondary swell wave height
4	VHM0_WW	Spectral significant wind wave height
5	VMDR	Mean wave direction from (Mdir)
6	VMDR_SW1	Mean primary swell wave direction from
7	VMDR_SW2	Mean secondary swell wave direction from
8	VMDR_WW	Mean wind wave direction from
9	VPED	Wave principal direction at spectral peak
10	VSDX	Stokes drift U
11	VSDY	Stokes drift V
12	VTM01_SW1	Spectral moments (0,1) primary swell wave period
13	VTM01_SW2	Spectral moments (0,1) secondary swell wave period
14	VTM01_WW	Spectral moments (0,1) wind wave period
15	VTM02	Spectral moments (0,2) wave period (Tm02)
16	VTM10	Spectral moments (-1,0) wave period (Tm-10)
17	VTPK	Wave period at spectral peak/peak period (Tp)
18	WIND_MAG	Wind magnitude
19	WIND_DIR	Wind direction
20-29	VHM0_LX	Lag components of spectral significant wave height (Hm0); VHM0_L1 for one hour lag, VHM0_L2 for two-hour lag, etc.
30-39	WIND_MAG_LX	Lag components of wind magnitude; WIND_MAG_L1 for one hour lag, WIND_MAG_L2 for two-hour lag, etc.

2.2. Field wave measurements / *Terenska mjerenja valova*

The measurements were made using the well-known DATAWELL Waverider DWR MKIII, which was anchored in cooperation with the Hydrographic Institute of the Republic of Croatia. The moored wave rider measures wave direction, wave height, and peak period. The measured data is stored on the internal data logger of the buoy, but also through the HF antenna connection on the buoy, the data is transmitted to the RX - C receiver on shore. The receiver is connected to a computer with a software package needed to collect and analyze the data. It is also equipped with GPS for positioning and tracking the buoy. The high-capacity batteries inside the Waverider ensure operation for up to one year without battery replacement.

Table 3 Names, geographic coordinates, and measurement periods wave buoys

Tablica 3. Imena, zemljopisne koordinate i razdoblja mjerenja plutača

Buoy	Name	Period	Latitude	Longitude
1	Rijeka	1.7.2009-30.6.2011	45.33° N	14.39° E
2	Split	1.11.2007-15.11.2008	43.49° N	13.17° E
3	Istra	1.11.2007-31.12.2008	44.74° N	16.47° E

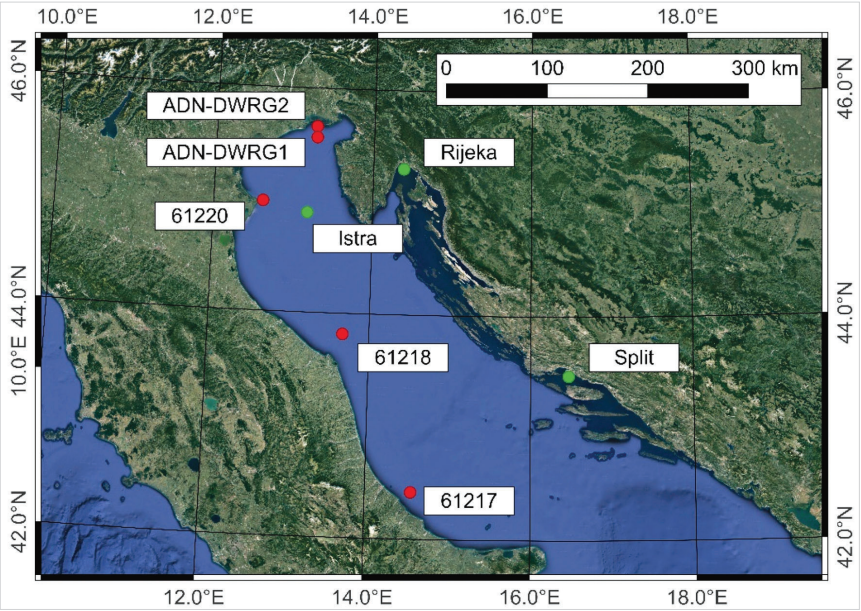

Figure 1 Position and names of wave buoys where measurements were conducted (green dots) and buoys used for MEDSEA validation presented in Table 1 (red dots)

Slika 1. Položaj i imena plutača vala gdje su izvršena mjerenja (zelene točke) i plutače upotrijebljene za MEDSEA validaciju koja je prikazana na Tablici 1 (crvene točke)

2.3. Machine learning models / *Modeli strojnoga učenja*

Three methods were used as correctors for the MEDSEA hindcast data at the Rijeka and Split sites. The MEDSEA hindcast for the Istra site showed a favorable performance metric, so no correction methods were applied to this dataset (described in more detail in Section 3.1.). Sections 2.3.1 and 2.3.2 describe the correction methods used in this study, ensemble regression, and artificial neural networks, respectively. Before the training procedure, the first 20% of the measured field data (described in Section 2.2.) were separated and labeled as test data set for testing the trained machine learning model (from November 1, 2007, to December 29, 2007, for Split and from June 1, 2009, to October 30, 2009, for Rijeka). The remaining data (80% of the total measured data) were separated for k-fold training of ensemble regression and artificial neural network models in 5 folds.

2.3.1. Ensemble regression (ER) - Least Squares Boosting Ensemble (LSBoost) / *Ansambl regresija (ER) – Boosting ansambl najmanjih kvadrata (LSBoost)*

Ensembles regression aggregates a set of trained weak learners (also called individual learners) to predict an ensemble response. The weak learner in this work is a decision tree, and the Least-squares boosting (LSBoost) method was used during cross-validation training of the ensemble regression model [40-42]. The algorithm first creates an initial model of the selected weak learner. Then at each time step, LSBoost fits a new weak learner to the current residuals, i.e., the difference between the observed response and the aggregated prediction of all learners created previously. Eventually, the aggregation of the weak learners should cause the residuals to converge. The LSBoost weighting uses the least-squares function as the loss function. In addition, Bayesian optimization was used to find the best possible hyperparameters values for the ensemble training: Number of learning cycles, learning rate, minimum leaf size and the maximum number of splits. The entire process was iterated 50 times to find the best hyperparameter combination. The optimization fitness objective was the mean squared error (described in Section 2.4). After training, the model was regularized using the Lasso algorithm, which finds an optimal set of weak learner weights and reduces overfitting to the data:

$$\sum_{n=1}^{N} w_n g\left(\left(\sum_{t=1}^{T}\alpha_t h_t\left(x_n\right)\right), y_n\right) + \lambda\sum_{t=1}^{T}\left|\alpha_t\right| \tag{2}$$

where λ is the lasso parameter, h_t is a weak learner in the ensemble trained on N observations with predictors x_n, responses y_n, and weights w_n. The λ value for regularization used in this work was 0.0008.

2.3.2. Artificial neural network / *Umjetna neuralna mreža*

A multi-layer feed-forward network was used to fit the function linking the input data (MEDSEA and ERA5, as described in section 2.1.) and the response data (measured field data, as described in section 2.2.) [22]. This is a common type of ANNs, where the output of each node in each layer is passed only to the following layer [43]. An ANN consists of input, hidden, and output nodes arranged in layers. Each input node is connected to multiple nodes, which together form the hidden layer or multiple connected hidden layers. Information is passed from the input nodes forward to nodes in the hidden layer:

$$h_j = f\left(a_j + \sum_{i=1}^{n} w_i x_i\right) \tag{3}$$

where x_i is the input variables, h_j is the responses of the hidden layer neuron, w_i is the weights, *aj* is the biases, and *f* is the activation function. Finally, the hidden layer was fully connected to the output layer, which consists of 1 node corresponding to the corrected significant wave height. The algorithm used the limited-memory Broyden-Fletcher-Goldfarb-Shanno quasi-Newton algorithm (LBFGS), where the mean squared error (*MSE*) was the optimization objective for training the weights and biases [44].

In addition, Bayesian optimization was used to find the best possible hyperparameter values for ANN training: activation function (relu, tanh, or sigmoid), lambda, number of layers, and layer sizes). The whole process was iterated 50 times to find the best hyperparameter combination. Lambda is a regularization strength term that counteracts the tendency of the training procedure to overfit the network to the training data. This is evident when statistical error metrics for the training and test datasets are drastically different [45].

The input and response data were preprocessed to efficiently train the ANNs. This normalization helped to avoid very small gradients and thus long training times. A common approach was used to normalize inputs and responses to fall within the range [–1, 1].

2.4. Statistical error metrics / *Statistička metrika pogrešaka*

To evaluate the prediction accuracy of the machine learning model and MEDSEA model, various statistical error metrics are used such as the coefficient of determination (R^2), normalized bias (*NBIAS*), the normalized root mean square error (*NRMSE*), mean absolute percentage error (*MAPE*), and scatter index (*SI*) as defined in Equations (5)–(9) respectively:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(P_i - O_i\right)^2}{\sum_{i=1}^{N}\left(P_i - \overline{P}\right)^2} \tag{5}$$

$$NBIAS = \frac{\overline{P} - \overline{O}}{\overline{O}} \tag{6}$$

$$NRMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(P_i - O_i\right)^2}{\sum_{i=1}^{N} O_i^2}} \tag{7}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{P_i - O_i}{P_i}\right| \tag{8}$$

$$SI = \sqrt{\frac{\sum_{i=1}^{N}\left[\left(P_i - \overline{P}\right) - \left(O_i - \overline{O}\right)\right]^2}{\sum_{i=1}^{N} O_i^2}} \tag{9}$$

where P_i is the ith prediction, O_i is the ith observation, $\overline{P}$ is the average prediction, and $\overline{O}$ is the average observation.

3. RESULTS AND DISCUSSION / *Rezultati i rasprava*

To determine if the MEDSEA modeled significant wave height data need to be corrected for the Rijeka, Istra, or Split sites, comparisons of modeled using MEDSEA and measured wave data are presented in Section 3.1. Based on the *NRMSE* and R^2 error metrics, the comparisons are presented and discussed whether a correction is needed. No resampling of the modeled MEDSEA was done with wave data since both have the same time step of 1 h. Then, machine learning correction methods (described in Sections 2.3.1. and 2.3.2.) are applied to the

modeled wave data to achieve a better fit to the measured data. The corrected significant wave height data are compared to measured wave data for Split and Rijeka in Sections 3.2.1. for ensemble regression and 3.2.2. for ANN model fit. Section 3.3 presents the improvements by adding lag components to the predictor set as described in Section 2.1. Finally, in Section 3.4. we evaluate the importance of the predictors. This is an inherent property of ensemble regression that can be easily extracted from a trained model.

3.1. Comparison of MEDSEA reanalysis modeled data and field wave measurements / *Usporedba MEDSEA reanalizom modeliranih podataka i stvarnih izmjerenih valova*

MEDSEA-modeled wave data showed good agreement with measurements near the Istra buoy, with an *NRMSE* value of 0.32 (Figure 2). This error is lower than the statistical metrics reported in the MEDSEA report (Table 1) [21] for other locations in the Adriatic Sea. Also, the modeled data showed low overprediction bias (2%). A similar *NRMSE* with the MEDSEA report values is to be expected since both the Istra buoy and the MEDSEA buoys (Figure 1) are not sheltered by islands that would reduce the fetch length below 30 km.

On the other hand, a substantial difference was found between the MEDSEA reanalysis modeled significant wave heights in comparison to the measured wave data for the locations of Split and Rijeka (Figure 2). The *NRMSE* increased to 0.52 and 0.74 for Split and Rijeka, respectively. This is 62% and 131% more than the *NRMSE* for Istra. These error metrics are also higher than the *NRMSE* values reported for other locations in the Adriatic (Table 1) [21]. The Split buoy wave measurements are moderately underpredicted (12%), while those at the Rijeka buoy are significantly underpredicted (35%). It is important to point out that the most likely reason for this discrepancy is the extremely complex orography surrounding the Rijeka and Split buoys, whose fetch length is less than 30 km. This is not the case for the Istra and MEDSEA buoys. This decrease in MEDSEA model accuracy in the case of Split and Rijeka is likely caused by unresolved topography by the wind and wave models and fetch accuracy limitations caused by the wave model resolution, as reported by Korres et al. [21]. This is all due to the eastern Adriatic Sea being enclosed basins near the coast with small fetch lengths dominated by wind waves.

Furthermore, the measured wave direction is spread mainly between the SW and SE directions, which is to be expected given the proximity of the buoys to the mainland in the NW direction (less than 1 km) (Figure 3 - right). Instead, the wave rose for the MEDSEA modeled data shows a strong presence of waves from the NE direction (Figure 3 - left). Although Korres et al. [21] did not compare the results of their modeled direction results to the measured wave directions, this preliminary observation shows that modeled wave directions for nearshore locations should be used with caution.

Overall, these results suggest that correction methods are not required for the station of Istra but are required for the stations of Rijeka and Split. The performance of these correction methods is presented in section 3.2.1. to 3.2.3.

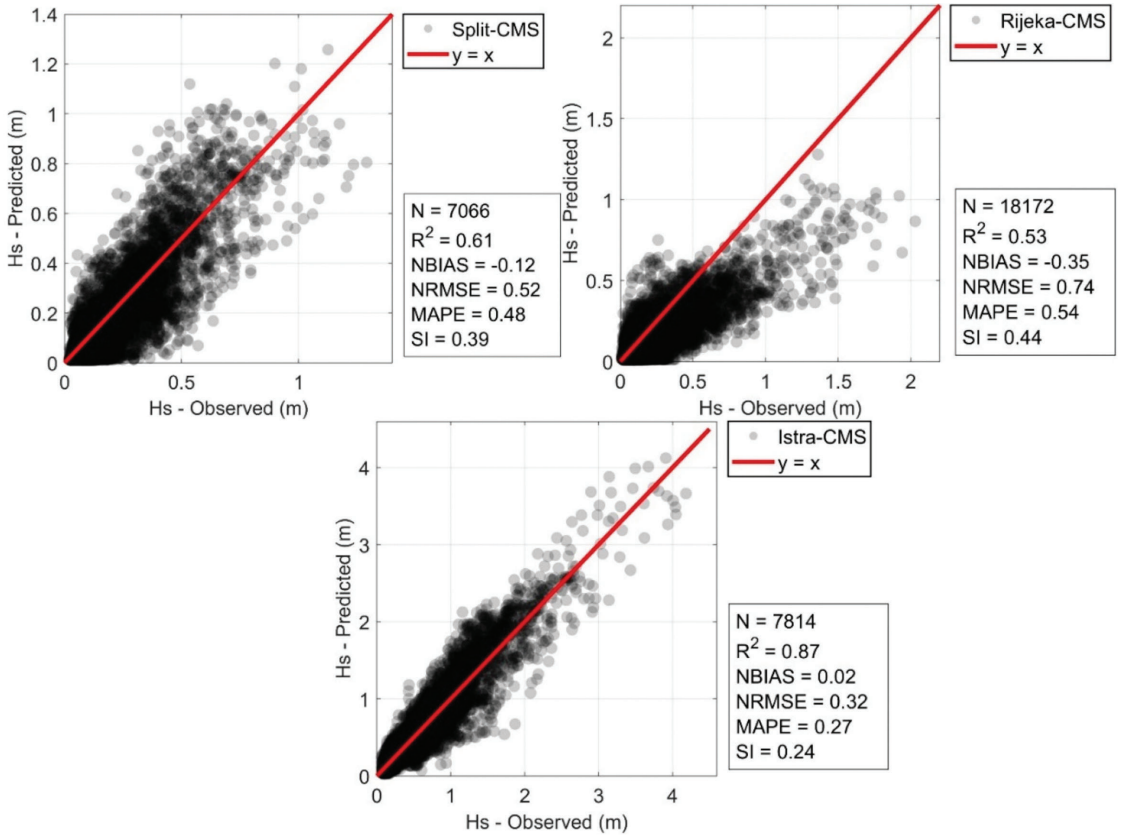

Figure 2 Correlation between MEDSEA reanalysis modeled (predicted) and measured (observed) significant wave heights for locations Split, Rijeka, and Istra

Slika 2. Korelacija između MEDSEA reanalizom modeliranih (predviđenih) i izmjerenih (promatranih) značajnih visina vala za lokacije Split, Rijeka i Istra

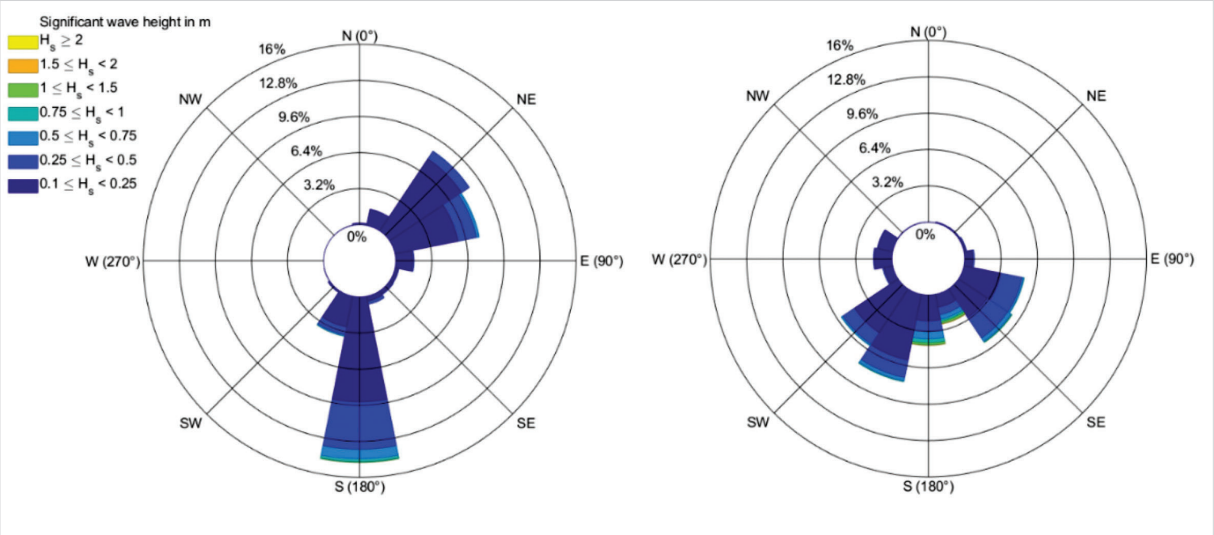

Figure 3 Wave rose of modeled wave heights (left) and measured wave heights (right) for the measuring period of 1.6.2009 - 1.6.2011 at the Rijeka buoy station

Slika 3. Ruža vjetrova modeliranih visina valova (lijevo) i izmjerenih valnih visina (desno) za mjerni period od 1. lipnja 2009. do 1. lipnja 2011. na lokaciji plutače Rijeka

3.2. Applying machine learning models to correct MEDSEA hindcast significant wave height data / *Primjena modela strojnoga učenja za ispravak MEDSEA podataka retrospektivno značajne visine vala*

3.2.1. Ensemble regression (ER) / *Regresija sklopa*

The statistical error metrics in Figure 4 show a moderate improvement of the initial MEDSEA hindcast with the ensemble regression correction method, with a stronger correlation coefficient and lower *NRMSE*, *NBIAS*, *MAPE*, and *SI*. The *NRMSE* are 0.53 and 0.48 for Rijeka and Split, respectively. This is a 29% and 8% reduction from the initial MEDSEA hindcast for Rijeka and Split, respectively. The *NBIAS* was reduced to a negligible value for both Rijeka and Split, with values of 0.01 and 0.04, respectively.

Table 4 summarizes the values of the hyperparameters resulting from the Bayesian estimation of the ensemble regression training procedure. While the number of learning cycles and the learning

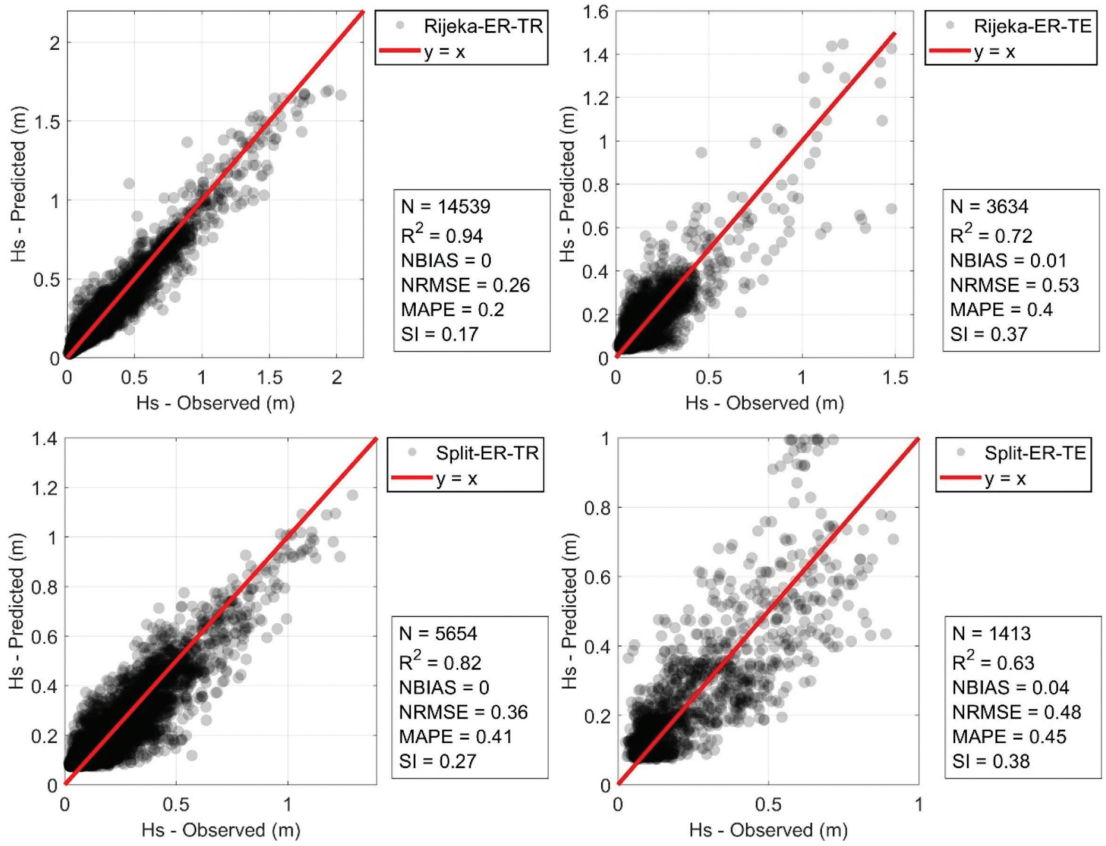

Figure 4 Statistical error metrics for evaluating MEDSEA hindcast wave heights corrected with ensemble regression (predicted) and measured (observed) wave heights for locations Split and Rijeka; TE – test data, TR – training data

Slika 4. Statistička metrika grešaka za evaluaciju MEDSEA prognoziranih valnih visina korigirano s ansambl regresijom (predviđeno) i izmjerene (promatrane) valne visine za lokacije Split i Rijeka; TE – podaci testa, TR – pokusni podaci

Table 4 Hyperparameters of the best-evaluated ensemble regression model after 5-fold cross-validation training

Tablica 4. Hiperparametri najbolje evaluiranih regresijskih modela prema peterostrukom validacijskom pokusu

	Name	Num. of learning cycles	Learn rate	Min. leaf size	Max. num. splits	Lambda
1	Rijeka	160	0.07019	8	11450	0.002
2	Split	138	0.10678	1	15	0.002

rate is comparable for the Rijeka and Split sites, the values for the minimum leaf size and the maximum number of splits are significantly larger for the Rijeka site. For both sites, the same lasso term lambda is used to prune the constructed ensemble of learning trees to avoid overfitting (the term λ in Eq. 2). Increasing the lasso term above this value rapidly degrades the predictive power of the ensemble regression for both the training and test sets (not shown in the manuscript for brevity), hence the value 0.002.

In summary, the statistical error metrics show the decent performance of the ensemble regression corrected MEDSEA hindcast of the significant wave height with a slight overestimation of the significant wave height.

3.2.2. Artificial neural networks (ANN) / *Umjetna neuralna mreža (ANN)*

In Figure 5, the statistical error metrics show a moderate improvement of the initial hindcast by MEDSEA with ANN as the correction method with higher correlation coefficients and lower NRMSE, NBIAS, MAPE, and SI. The NRMSE for the hindcast corrected with ANN is 0.55 and 0.46 for Rijeka and Split, respectively. This represents a 26% and 12% decrease in *NRMSE*, respectively, compared to the initial MEDSEA hindcast. The NBIAS was completely removed from the hindcast for the Rijeka site, while it was reduced to a negligible value of 0.02 for the Split-site.

Table 5 shows the hyperparameter values of ANN, which performed best in Bayesian optimization after 5-fold cross-validation on the training data. The sigmoid activation function is used for both Rijeka and Split. The hidden layers are both shallow with only 1 and 2 layers for Split and Rijeka, respectively, and narrow with only a maximum of 17 knots (Rijeka).

Overall, the results of the statistical analysis show a good performance of the corrected hindcast of the significant wave height with ANN with a slight overestimation of the significant wave height for the Split site.

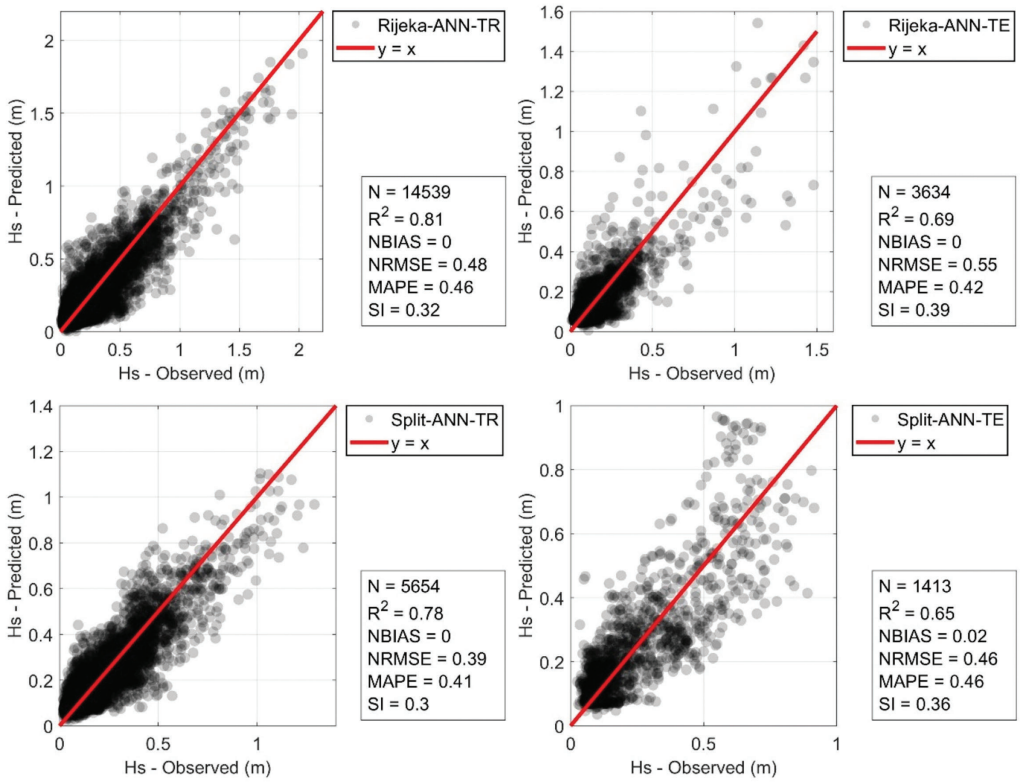

Figure 5 Statistical error metrics for evaluating MEDSEA hindcast wave heights corrected with ANN and measured (observed) wave heights for locations Split and Rijeka; TE – test data, TR – training data

Slika 5. Statistička metrika grešaka za evaluaciju MEDSEA prognoziranih valnih visina korigirano s ANN i izmjerene (promatrane) visine vala za lokacije Split i Rijeka; TE – podaci testa, TR – pokusni podaci

Table 5 Hyperparameters of the best evaluated ANN after 5-fold cross-validation training

Tablica 5. Hiperparametri najbolje evaluiranih umjetnih neuralnih mreža prema peterostrukom validacijskom pokusu

	Name	Activation function	Lambda	Layer Sizes
1	Rijeka	sigmoid	6.2882×10^{-6}	17 15
2	Split	sigmoid	1.8652×10^{-6}	4

3.2.3. Impact of lag components and time series comparison / *Utjecaj komponenti s vremenskim pomakom i usporedba vremenskih serija*

Interestingly, there is little difference in the statistical error metrics due to the addition of lag components of wind magnitude from ERA5 and significant wave height from MEDSEA to the predictor set (described in section 2.1) (Table 6). Error metrics varied in the range of ±8%, while some metrics did not change with the inclusion of lag components. Only the *NBIAS* changed significantly, as its values were already low before the inclusion of the lag components. Therefore, no significant difference in corrected model performance was found between models with and without lag components. However, the cross-correlation for the Split site calculated between the signals of the measured significant wave height, VHM0, and the lag components of the ERA5 wind magnitude, WIND_MAG, shows that the 1-hour lag component has the highest correlation with the measured VHM0 (0.82). The 0-hour lag component has a slightly lower cross-correlation of 0.81, while the higher lag components steadily decrease in value from the 1-hour lag component. This trend suggests a diffused cross-correlation without a dominant lag component that would have significant explanatory power. The Split location is examined here for cross-validation because in Section 3.2.4, the WIND_MAG variable shows the highest predictor significance of all predictors used.

Figure 6 shows a time series of measured, MEDSEA-modeled, and MEDSEA-corrected significant wave heights at the Rijeka site. The MEDSEA reanalysis model does not respond quickly enough to the increasingly significant wave height on October 21 and eventually underestimates by 0.6 m the largest wave height observed on October 22. The corrected models (ER and ANN) can correct the underestimation, with the ANN model overestimating the largest wave height by up to 0.2 m, but was unable to accelerate the increase in wave height at the beginning of the wave event. The corrected models with the lag components (ER-L and ANN-L) showed no significant improvement over the corrected models without lag components (ER and ANN).

Table 6 Statistical error metrics of test data for the correction methods that include lag components for wind magnitude from ERA5 and significant wave height from MEDSEA, as described in section 2.1; the percentage is relative to the error metrics of correction methods excluding the lag components (shown in sections 3.2.1. and 3.2.2)

Tablica 6. Statistička metrika grešaka testnih podataka za metode korekcije koja uključuje komponente s vremenskim pomakom za magnitude vjetra od ERA5 i značajne visine vala od MEDESEA, kao što je opisano u sekciji 2.1; postotak je relativan u odnosu na metriku metoda korekcije isključujući komponente zaostajanja (prikazane u sekcijama 3.2. i 3.2.2)

	Name	R2 (%)	NRMSE (%)	NBIAS (%)	MAPE (%)	SI (%)
1	Rijeka-L-ER	0.73 (+1%)	0.52 (-2%)	0.01 (+0%)	0.38 (-5%)	0.37 (+0%)
2	Split-L-ER	0.59 (-6%)	0.50 (+4%)	0.03 (-25%)	0.46 (-4%)	0.39 (+2%)
3	Rijeka-L-ANN	0.69 (+0%)	0.55 (+0%)	-0.03 (∞%)	0.37 (-8%)	0.39 (+2%)
4	Split-L-ANN	0.66 (+2%)	0.46 (+0%)	0.02 (+0%)	0.47 (+2%)	0.36 (+0%)

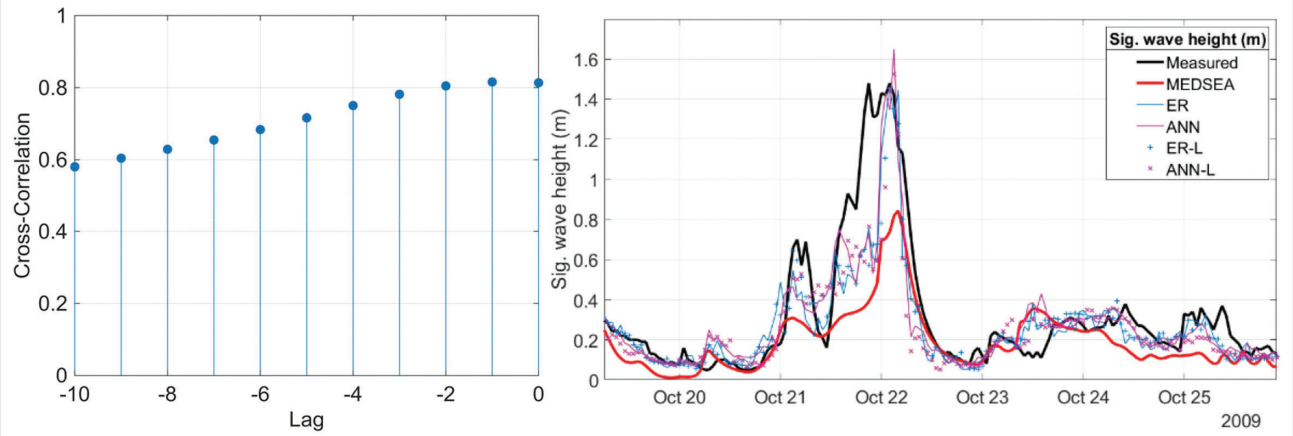

Figure 6 (left) Cross-correlation between the measured significant wave height, VHM0, and the lag components of the ERA5 wind magnitude WIND_MAG for the location of Split; right) Time series of measured sig. wave heights, MEDSEA modeled, ensemble regression corrected without (ER) and with lag components (ER-L), ANN corrected without (ANN) and with lag components (ANN-L)

Slika 6. (lijevo) Unakrsna korelacija između izmjerene značajne visine vala, VHMO i komponenti s vremenskim pomakom ERA5 magnitude vjetra WIND MAG za lokaciju Split; (desno) Vremenske serije izmjerenih visina valova modeliranih pomoću MEDSEA, ansambl regresija korigirana bez (ER) i s komponentama s vremenskim pomakom (ER-L), ANN korigirano bez (ANN) i s komponentama s vremenskim pomakom (ANN-L)

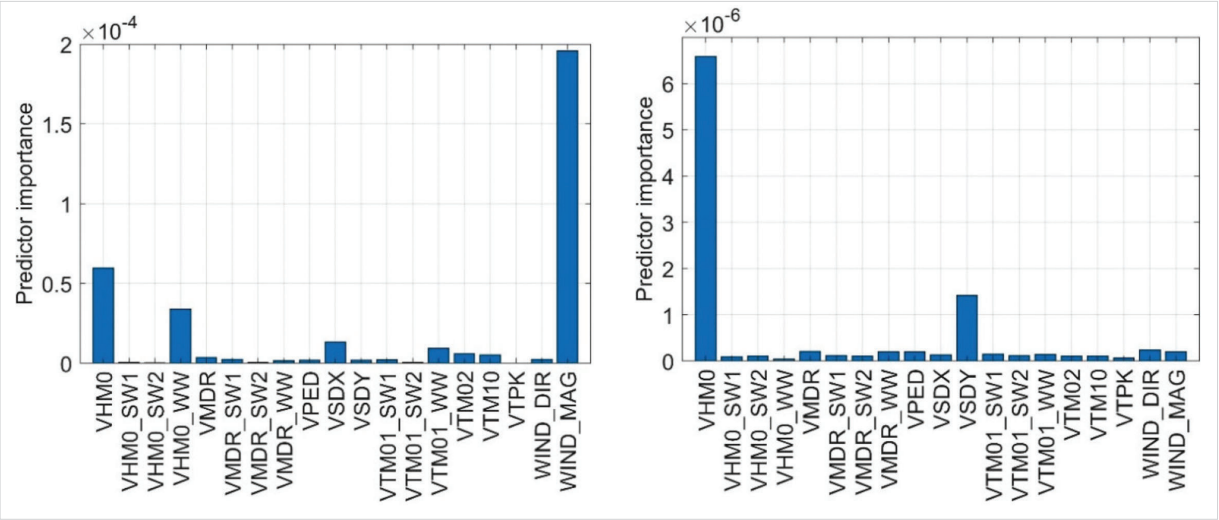

Figure 7 Predictor importance in ensemble regression (name designations are described in Table 2) for Split (left) and Rijeka (right)
Slika 7. Važnost pojedinog prediktora pri ansambl regresiji (imena parametara opisana su u Tablici 2) za Split i Rijeku (desno)

3.2.4. Predictor importance / *Važnost prediktora*

Predictor importance is calculated by summing the changes in node risk due to splits on each predictor variable in a regression tree or ensemble of regression trees and then dividing the sum by the number of branch nodes. The change in node risk is the difference between the risk of the parent node and the combined risk for the two child nodes. Node risk is defined as the mean squared error of the node weighted by the node probability. In this way, the relative importance of the predictors can be extracted from trained ensemble regression models (Figure 7).

Figure 7 shows that the wind magnitude from ERA5 (WIND_MAG) is the most important predictor for Split, but significant wave height from MEDSEA (VHM0) is the most important predictor for Rijeka. However, significant wave height from MEDSEA (VHM0) still has a significant effect on model performance in Split, albeit one-four times smaller than wind magnitude. Other predictors have a smaller effect on correction model performance. There could be several reasons for this difference. Both sites are located in sheltered basins where the wind is the dominant wave generator. Therefore, the wind strength modeled by ERA5 in Rijeka could be underestimated due to the complex orography, similar to the underestimation of wind magnitude underestimation in the Ligurian Sea in Italy [39].

4. CONCLUSION / *Zaključak*

This paper presents a methodology for extending hindcast wave data to sparsely measured locations based on machine learning models and reanalysis data. The advantage of a machine learning model (ANN, ensemble regression, etc.) is that no location-specific data are needed for hindcasting wave parameters. Only publicly available global or regional reanalysis model data could be used as input for training and eventually hindcasting wave parameters or filling gaps in existing wave measurements. This is particularly important in locations sheltered by complex topography, as nested wave models are typically required to properly represent wave processes from the open ocean to sheltered basins. These nested numerical models require more time to set up and compute. However, unlike machine learning models, numerical models are constrained by physical properties. Therefore, machine learning models should be used with caution. Because the present technique efficiently improves local nearshore waves from MEDSEA with low computational cost to better reflect measured data, it could be applied in locations with sparse wave observations to augment measured wave data.

This work has shown that the machine learning hindcast did not improve the Split initial MEDSEA hindcast as well as the Rijeka hindcast. The *NRMSE* improvement for Rijeka is 29% for the test data (Rijeka-ER-TE), compared to the smaller improvement for Split of 12% (Split-ANN-TE). This smaller improvement for Split may be because the initial wave height hindcasts had higher accuracy, making it difficult for the machine learning models to further improve accuracy. Furthermore, the machine learning models reduced the biases in the MEDSEA hindcasts to negligible levels for both Split and Rijeka (*NBIAS* < 0.03). Nevertheless, the presented machine learning method could not improve the hindcasts for Split and Rijeka to the level of MEDSEA hindcasts for Istra (Figure 2) or other buoys in the open sea of the Adriatic (Table 1).

Interestingly, the results showed that wind duration (via the wind magnitude lag components from -1 h to -10 h) and wave height history (via wave height lag components from -1 h to -10 h) as input data did not significantly improve the performance of the wave height hindcast. With a marginal improvement in statistical error metrics, this is considered a negligible improvement over previously established machine learning models without lag components. This however is not aligned with the observation done with the cross-correlation analysis between the measured significant wave heights and the ERA5 wind magnitudes, where the 1-hour lag component showed the highest cross-correlation. These results also do not agree with those of Peres et al. [8], who found that including up to 18 hours of wind lag components improved wave height hindcast with ANN. However, Peres et al. [8] examined the performance of ANN in the western Mediterranean, i.e., not in extremely sheltered areas with small fetch lengths. MEDSEA greatly miscalculates the wave directions, as can be seen in Figure 3.

Therefore, MEDSEA wave direction should be used with caution in sheltered areas because the spatial resolution of the model is not sufficient to accurately resolve the wave processes.

In the future, other methods should be explored, such as introducing a weight distribution into the machine learning objective function to increase the penalty for errors at extreme wave heights. This should improve the hindcast for extreme wave heights at the expense of reduced accuracy at lower wave heights. In addition, a random separation between the training and test sets was performed for this study. A carefully performed data separation and curation could lead to even further increases in accuracy. In addition, a hybrid approach of nested numerical wave modeling and machine learning is also a viable option. The physically modeled wave results would serve as an extension of the measured wave data, which would incur higher computational costs to obtain a longer time series on which to train the machine learning models. Finally, measured meteorological data could be used instead of or in addition to the reanalysis data to reconstruct measured wave heights.

Author Contributions: Conceptualization, D. B.; Methodology, D. B. and D. C.; Software, D. B., and T. B.; Validation, D. C.; Formal Analysis, D. B., T. B. and T. K.; Investigation, D. B., T. B., T. K.; Resources, T. B.; Data Curation, T. B.; Writing – Original Draft Preparation, D. B.; Writing – Review & Editing, D. B., T. B., T. K., and D. C.; Visualization, D. B.; Supervision, D. C.; Project Administration, D. C., and T. B.; Funding Acquisition, D. C.

Funding: This work has been fully supported by the "Research Cooperability" Program of the Croatian Science Foundation funded by the European Union from the European Social Fund under the Operational Programme Efficient Human Resources 2014-2020.

Conflict of interest: None.

REFERENCES / *Literatura*

[1] Nitsure, S. P., Londhe, S. N. & Khare, K. C. (2012). Wave forecasts using wind information and genetic programming. *Ocean Engineering*, *54*, 61-69. https://doi.org/10.1016/j.oceaneng.2012.07.017

[2] Goda, Y. (1985). *Random Seas and Design of Maritime Structure*. The University of Tokyo Press.

[3] Bosom, E. & Jimenez, J. A. (2011). Probabilistic coastal vulnerability assessment to storms at regional scale - application to Catalan beaches (NW Mediterranean). *Natural Hazards and Earth System Sciences*, *11*(2), 475-484. https://doi.org/10.5194/nhess-11-475-2011

[4] Camus, P., Mendez, F. J. & Medina, R. (2011). A hybrid efficient method to downscale wave climate to coastal areas. *Coastal Engineering*, *58*(9), 851-862. https://doi.org/10.1016/j.coastaleng.2011.05.007

[5] Vannucchi, V., Taddei, S., Capecchi, V., Bendoni, M. & Brandini, C. (2021). Dynamical Downscaling of ERA5 Data on the North-Western Mediterranean Sea: From Atmosphere to High-Resolution Coastal Wave Climate. *Journal of Marine Science and Engineering*, *9*(2). https://doi.org/10.3390/jmse9020208

[6] Masselink, G., Kroon, A. & Davidson-Arnott, R. G. D. (2006). Morphodynamics of intertidal bars in wave-dominated coastal settings - A review. *Geomorphology*, *73*(1-2), 33-49. https://doi.org/10.1016/j.geomorph.2005.06.007

[7] Bellotti, G., Franco, L. & Cecioni, C. (2021). Regional Downscaling of Copernicus ERA5 Wave Data for Coastal Engineering Activities and Operational Coastal Services. *Water*, *13*(6). https://doi.org/10.3390/w13060859

[8] Peres, D. J., Iuppa, C., Cavallaro, L., Cancelliere, A. & Foti, E. (2015). Significant wave height record extension by neural networks and reanalysis wind data. *Ocean Modelling*, *94*, 128-140. https://doi.org/10.1016/j.ocemod.2015.08.002

[9] Copernicus Marine Service (CMEMS) In Situ TAC. (2022). *Dashboard*. Retrieved 27.04.2022. from http://www.marineinsitu.eu/dashboard/

[10] Pomaro, A., Cavaleri, L., Papa, A. & Lionello, P. (2018). 39 years of directional wave recorded data and relative problems, climatological implications and use. *Scientific Data*, *5*. https://doi.org/10.1038/sdata.2018.139

[11] World Meteorological Organization. (2018). *Guide to Wave Analysis and Forecasting* (Vol. WMO-No. 702).

[12] Goda, Y. (2003). Revisiting Wilson's formulas for simplified wind-wave prediction. *Journal of waterway port coastal and ocean engineering-asce*, *129*(2), 93-95. https://doi.org/10.1061/(Asce)0733-950x(2003)129:2(93)

[13] WAMDI Group. (1988). The WAM Model—A Third Generation Ocean Wave Prediction Model. *J. Phys. Oceanogr.*, *18*(2), 1775–1810. https://doi.org/doi:10.1175/1520-0485(1988)018<1775:TWMTGO>2.0.CO

[14] Booij, N., Ris, R. C. & Holthuijsen, L. H. (1999). A third-generation wave model for coastal regions - 1. Model description and validation. *Journal of Geophysical Research-Oceans*, *104*(C4), 7649-7666. https://doi.org/Doi 10.1029/98jc02622

[15] Kim, S., Tom, T. H. A., Takeda, M. & Mase, H. (2021). A framework for transformation to nearshore wave from global wave data using machine learning techniques: Validation at the Port of Hitachinaka, Japan. *Ocean Engineering*, *221*. https://doi.org/10.1016/j.oceaneng.2020.108516

[16] Feng, X. & Chen, X. (2021). Feasibility of ERA5 reanalysis wind dataset on wave simulation for the western inner-shelf of Yellow Sea. *Ocean Engineering*, *236*. https://doi.org/10.1016/j.oceaneng.2021.109413

[17] Reanalysis, A. (2010). *Ocean Reanalyses Table*. Retrieved 29.4.2022. from https://reanalyses.org/observations/ocean-reanalyses-table

[18] Smith, C. A., Compo, G. P. & Hooper, D. K. (2014). Web-Based Reanalysis Intercomparison Tools (WRIT) for Analysis and Comparison of Reanalyses and Other Datasets. *Bulletin of the American Meteorological Society*, *95*(11), 1671-1678. https://doi.org/10.1175/Bams-D-13-00192.1

[19] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Munoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., . . . Thepaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. https://doi.org/10.1002/qj.3803

[20] Law-Chune, S., Aouf, L., Dalphinet, A., Levier, B., Drillet, Y. & Drevillon, M. (2021). WAVERYS: a CMEMS global wave reanalysis during the altimetry period. *Ocean Dynamics*, *71*(3), 357-378. https://doi.org/10.1007/s10236-020-01433-w

[21] Korres, G., Ravdas, M. & Zacharioudaki, A. (2019). *Mediterranean Sea Waves Hindcast (CMEMS MED-Waves) [Data set]*. https://doi.org/https://doi.org/10.25423/CMCC/MEDSEA_HINDCAST_WAV_006_012

[22] Haykin, S. (2010). Neural networks: a comprehensive foundation. 1999. *Mc Millan, New Jersey*, 1-24.

[23] Berbić, J., Ocvirk, E., Carević, D. & Lončar, G. (2017). Application of neural networks and support vector machine for significant wave height prediction. *Oceanologia*, *59*(3), 331-349. https://doi.org/10.1016/j.oceano.2017.03.007

[24] Elbisy, M. S. & Elbisy, A. M. S. (2021). Prediction of significant wave height by artificial neural networks and multiple additive regression trees. *Ocean Engineering*, *230*. https://doi.org/10.1016/j.oceaneng.2021.109077

[25] Passarella, M., Goldstein, E. B., De Muro, S. & Coco, G. (2018). The use of genetic programming to develop a predictor of swash excursion on sandy beaches. *Natural Hazards and Earth System Sciences*, *18*(2), 599-611. https://doi.org/10.5194/nhess-18-599-2018

[26] van Maanen, B., Coco, G., Bryan, K. R. & Ruessink, B. G. (2010). The use of artificial neural networks to analyze and predict alongshore sediment transport. *Nonlinear Processes in Geophysics*, *17*(5), 395-404. https://doi.org/10.5194/npg-17-395-2010

[27] Goldstein, E. B., Coco, G. & Plant, N. G. (2019). A review of machine learning applications to coastal sediment transport and morphodynamics. *Earth-Science Reviews*, *194*, 97-108. https://doi.org/10.1016/j.earscirev.2019.04.022

[28] Bujak, D., Bogovac, T., Carević, D., Ilic, S. & Lončar, G. (2021). Application of Artificial Neural Networks to Predict Beach Nourishment Volume Requirements. *Journal of Marine Science and Engineering*, *9*(8). https://doi.org/10.3390/jmse9080786

[29] Mahjoobi, J., Etemad-Shahidi, A. & Kazeminezhad, M. H. (2008). Hindcasting of wave parameters using different soft computing methods. *Applied Ocean Research*, *30*(1), 28-36. https://doi.org/10.1016/j.apor.2008.03.002

[30] Londhe, S. N. (2008). Soft computing approach for real-time estimation of missing wave heights. *Ocean Engineering*, *35*(11-12), 1080-1089. https://doi.org/10.1016/j.oceaneng.2008.05.003

[31] Alexandre, E., Cuadra, L., Nieto-Borge, J. C., Candil-García, G., del Pino, M. & Salcedo-Sanz, S. (2015). A hybrid genetic algorithm—extreme learning machine approach for accurate significant wave height reconstruction. *Ocean Modelling*, *92*, 115-123. https://doi.org/10.1016/j.ocemod.2015.06.010

[32] Komen, G. J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S. & Janssen, P. (1994). *Dynamics and modelling of ocean waves*. Cambridge University Press.

[33] Janssen, P. A. E. M. (1989). Wave-Induced Stress and the Drag of Air Flow over Sea Waves. *Journal of Physical Oceanography*, *19*(6), 745-754. https://doi.org/10.1175/1520-0485(1989)019<0745:Wisatd>2.0.Co;2

[34] Janssen, P. (1991). Quasi-linear Theory of Wind-Wave Generation Applied to Wave Forecasting. *Journal of Physical Oceanography*, *21*, 1631-1642.

[35] Hasselmann, K. (1973). ON THE SPECTRAL DISSIPATION OF OCEAN WAVES DUE TO WHITE CAPPING. *Boundary-Layer Meteorology*, *6*, 107-127.

[36] Weatherall, P., Marks, K. M., Jakobsson, M., Schmitt, T., Tani, S., Arndt, J. E., Rovere, M., Chayes, D., Ferrini, V., & Wigley, R. (2015). A new digital bathymetric model of the world's oceans. *Earth and Space Science*, *2*(8), 331-345. https://doi.org/10.1002/2015ea000107

[37] Lionello, P., Gunther, H. & Janssen, P. A. E. M. (1992). Assimilation of Altimeter Data in a Global 3rd-Generation Wave Model. *Journal of Geophysical Research-Oceans*, *97*(C9), 14453-14474. https://doi.org/Doi 10.1029/92jc01055

[38] Bujak, D., Loncar, G., Carevic, D. & Kulic, T. (2023). The Feasibility of the ERA5 Forced Numerical Wave Model in Fetch-Limited Basins. *Journal of Marine Science and Engineering*, *11*(1). https://doi.org/10.3390/jmse11010059

[39] Davison, S., Benetazzo, A., Barbariol, F., Cavaleri, L. & Pezzutto, P. (2019). *Assessment of ERA5 winds in the Mediterranean Sea*.

[40] Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/Doi 10.1023/A:1010933404324

[41] Hastie, T., Tibshirani, R. & J., F. (2008). *The Elements of Statistical Learning* (second edition ed.). Springer.

[42] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232. https://doi.org/DOI 10.1214/aos/1013203451

[43] van Gent, M. R. A., van den Boogaard, H. F. P., Pozueta, B., & Medina, J. R. (2007). Neural network modelling of wave overtopping at coastal structures. *Coastal Engineering*, *54*(8), 586-593. https://doi.org/10.1016/j.coastaleng.2006.12.001

[44] Nocedal, J., & Wright, S. J. (2006). *Numerical Optimization* (2nd ed. ed.). Springer.

[45] Medina, J. R., Garrido, J., Gómez-Martín, E., & Vidal, C. (2003). Armour damage analysis using neural networks. Proceedings of Coastal Structures 2003, Portland, USA.