Applied Ocean Research 153 (2024) 104273

Contents lists available at ScienceDirect

Applied Ocean Research

journal homepage: www.elsevier.com/locate/apor

Fuzzy-based ensemble methodology for accurate long-term prediction and interpretation of extreme significant wave height events

C. Peláez-Rodríguez [a], J. Pérez-Aracil [a], A.M. Gómez-Orellana [b], D. Guijo-Rubio [a], V.M. Vargas [b,*], P.A. Gutiérrez [b], C. Hervás-Martínez [b], S. Salcedo-Sanz [a]

[a] Department of Signal Processing and Communications, Universidad de Alcalá, Alcalá de Henares, 28805, Spain
[b] Department of Computer Science and Numerical Analysis, Universidad de Córdoba, Córdoba, 14041, Spain

ARTICLE INFO

Keywords:
Extreme significant wave height
Energy flux
Ensemble models
Long-term prediction
Explainable artificial intelligence

ABSTRACT

Providing an accurate prediction of Significant Wave Height (SWH), and specially of extreme SWH events, is crucial for coastal engineering activities and holds major implications in several sectors as offshore renewable energy. With the aim of overcoming the challenge of skewness and imbalance associated with the prediction of these extreme SWH events, a fuzzy-based cascade ensemble of regression models is proposed. This methodology allows to remarkably improve the predictive performance on the extreme SWH values, by using different models specialised in different ranges on the target domain. The method's explainability is enhanced by analysing the contribution of each model, aiding in identifying those predictor variables more characteristic for the detection of extreme SWH events. The methodology has been validated tackling a long-term SWH prediction problem, considering two case studies over the southwest coast of the United States of America. Both reanalysis data, providing information on various meteorological factors, and SWH measurements, obtained from the nearby stations and the station under examination, have been considered. The goodness of the proposed approach has been validated by comparing its performance against several machine learning and deep learning regression techniques, leading to the conclusion that fuzzy ensemble models perform much better in the prediction of extreme events, at the cost of a slight deterioration in the rest of the samples. The study contributes to advancing the SWH prediction field, specially, to understanding the behaviour behind extreme SWH events, critical for various sectors reliant on oceanic conditions.

1. Introduction

Surface wind waves represent a stochastic phenomenon with the potential to significantly impact various fields including science, logistics, and technology. These waves can affect ship trajectory and speed, leading to hull resonance and fracture risks. Moreover, they pose threats to ports, underwater structures, and coastal defences (Feng et al., 2022). On the other hand, the increasing interest in offshore renewable energy, particularly in wave energy harvesting technologies like Wave Energy Converters (WECs), has underscored the importance of understanding wave conditions and predicting Significant Wave Height (SWH), which represent one of the most important parameters in this regard (Falcao, 2010; Guijo-Rubio et al., 2020). Furthermore, due to the intermittent and stochastic nature of waves, precise SWH prediction presents a fundamental challenge with far-reaching implications for WECs, marine-related industries, such as optimising wave farms or shipping routes, and for evaluating extreme wave loads on marine structures such as wind turbines (Zilong and Wei, 2022) or cross-bridges (Ti et al., 2020). Recent advancements in this area have further emphasised the necessity of accurate SWH prediction for effective management of offshore activities.

SWH prediction models generally fall into two categories: data-driven (Mudronja et al., 2017) or physical-driven (Ibarra-Berastegi et al., 2015) frameworks. While numerical models based on energy transfer equations offer effective results across large-scale space and time ranges, they often require intensive computing resources and time (Güner et al., 2013). Alternatively, statistical approaches and Artificial Intelligence (AI) techniques leverage extensive datasets to make predictions by identifying potential relationships and dependencies between variables. With the proliferation of buoy stations providing real-time SWH data and historical events, soft computing techniques, including Machine Learning (ML) and Deep Learning (DL), are emerging as valuable tools in this field for a range of tasks, such

* Corresponding author.
E-mail address: vvargas@uco.es (V.M. Vargas).

https://doi.org/10.1016/j.apor.2024.104273
Received 30 April 2024; Received in revised form 27 September 2024; Accepted 11 October 2024

as SWH segmentation (Durán-Rosal et al., 2017) or SWH reconstruction (Guijo-Rubio et al., 2023), among others. Hence, these techniques hold promise for improving SWH prediction accuracy and efficiency, as evidenced by recent research efforts in the domain Shamshirband et al. (2020), Gómez-Orellana et al. (2022), Zilong et al. (2022), Afzal et al. (2023), Ding et al. (2023), Minuzzi and Farina (2024), Gómez-Orellana et al. (2024), Abbas et al. (2024).

In this context, extreme SWH plays a critical role in coastal engineering activities and have significant geophysical implications. Consequently, studying, observing, and predicting these waves from a few hours to a few days in advance is crucial for daily marine tasks (Dysthe et al., 2008), and remains a challenging problem due to the extremely complex, non-stationary, non-linear, and uncertain nature of their physical generation process (Dixit and Londhe, 2016). Within the oceanographic community, extreme SWH events are characterised by surface gravity waves with heights significantly exceeding expected values for the prevailing sea state (Dysthe et al., 2008). Predicting these extreme SWH events has become an important area of research in recent years. For instance, in Dixit and Londhe (2016), a hybrid neuro wavelet technique is proposed for predicting extreme SWH during major hurricane events, with an anticipation from +12-h to +36-h. In Rueda et al. (2016), authors present a classification of weather patterns to statistically downscale daily SWH maxima to a local area of interest, demonstrating the model's ability to reproduce different time scales. Additionally, in Petrov et al. (2013), maximum entropy is introduced as a powerful tool for predicting extreme SWH, comparing its performance with models within the extreme value theory framework such as the generalised Pareto distribution and the generalised extreme values distribution. These studies highlight ongoing efforts to improve our understanding and prediction of extreme SWH events, with implications for various sectors reliant on oceanic conditions.

In this work, we focus on predicting extreme SWH, which poses challenges associated with databases characterised by high skewness and imbalance. This imbalance arises because instances with extreme SWHs often represent a minimal percentage of the overall dataset. To address this issue, predominant strategies can be categorised into two main groups: preprocessing techniques and ensemble methodologies. Preprocessing techniques like undersampling and oversampling are commonly employed to balance data, as discussed in Batista et al. (2004). On the contrary, ensemble methodologies involve a decision-making process that combines individual learning algorithms and their outputs to achieve the most accurate result. While some strategies related to ensemble learning can be found in the literature for addressing SWH prediction problems (Kumar et al., 2018), there is currently a lack of extensive research in the field of extreme SWH prediction. This area holds promise for future investigation to improve the accuracy and reliability of extreme SWH predictions. Ensemble learning has been applied to a variety of prediction problems related to meteorological variables (Ren et al., 2016; Chen et al., 2018; Farahbod et al., 2022; Liu et al., 2021). In this field, it is noteworthy to mention the fuzzy ensemble methods, which have been also explored in recent literature (Gao et al., 2022; Peláez-Rodríguez et al., 2023; Sideratos et al., 2020; Prado et al., 2020).

Besides, in the field of extreme weather events prediction, understanding the factors that contribute to their occurrence is essential for effective analysis. Therefore, alongside obtaining accurate methods, the ability to explain the underlying mechanisms driving these events is key. This aspect is known as eXplainable AI (XAI). XAI refers to the ability of AI systems to provide transparent explanations of their predictions and decision-making processes (Arrieta et al., 2020). By integrating XAI techniques into meteorological models and prediction systems, researchers and meteorologists can gain a better understanding of the underlying factors that contribute to weather patterns. The combination of explainability with accuracy in the design of new approaches, not only improve the accuracy of predictions, but also enhance trust and confidence in AI systems, allowing for more effective decision-making and response to weather-related events (Toms et al., 2020). Several solutions appear in the recent literature associated with the application of XAI in the field of renewable energy, meteorology and climate science: (Ilic et al., 2021; Akhlaghi et al., 2021; Iong et al., 2022; Gao and Wang, 2022; Sushanth et al., 2023; Gómez-Orellana et al., 2023; Peláez-Rodríguez et al., 2024). However, limited literature exists on methods specifically tailored for explaining predictions of extreme SWH events (Hansom et al., 2015; Samayam et al., 2017).

In this paper, we propose the application of a fuzzy-based ensemble methodology for accurate and explainable prediction of extreme SWH events. This methodology was first introduced by the authors in Peláez-Rodríguez et al. (2023) for improving extreme wind speeds prediction, and has been enhanced by various innovative contributions, significantly improving the accuracy and interpretability in predicting extreme events. Specifically, the proposed algorithm consists in a fuzzy-based cascade ensemble of regression models, where each model is focused on a specific part of the target domain. In this way, the explainability of the method is addressed by analysing the contributions of the specific models to the ensemble prediction, observing the distinct distributions of the various data subsets used for training the individual models, and identifying which predictor variables are characteristic of extreme SWH occurrences. This process enables the identification of disparities between extreme and non-extreme SWH events, further enhancing the interpretability of the method.

The proposed methodology is employed to tacked the long-term SWH prediction task, considering two distinct time prediction horizons: +24-h and +36-h. These prediction horizons are appropriately selected due to their ability to provide enough anticipation for crucial marine operations, as aforementioned. Specifically, the prediction is conducted for two stations located at the southwest coast of the United States of the America (USA). For this, reanalysis data providing information on various meteorological factors, such as air temperature or wind speed, are used along with SWH measurements obtained from the nearby stations and the station under examination. Note that both reanalysis variables and SWH measurements are included at the 0-h time instant. In addition to the long-term SWH prediction, an analysis of the extreme SWH events is also carried out for both stations, with the goal of gaining insights into the variables with the most influence on them.

Summarising, the contributions of this work are the following:

1. Development of a fuzzy-based ensemble methodology for the long-term SWH prediction. This methodology encompasses different models of the same regressor as components, each trained with data belonging to a specific range of SWH.
2. Design and application of the proposed methodology to two different case studies, in which the prediction is carried out considering two time prediction horizons: +24-h and +36-h.
3. Interpretation and analysis of the prediction carried out for extreme SWH events. The ranges of the predictive variables belonging to the individual models that contributes the most to obtain the final prediction are analysed, providing information on which variables are of most significance in the prediction of SWH extreme events.
4. Comparison of the proposed approach against several state-of-the-art ML and DL methods using numerous performance metrics, such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Additionally, the analysis of extreme SWH involves two variants of MAE and RMSE, known as Extreme Events MAE (EEMAE) and Extreme Events RMSE (EERMSE), along with the inclusion of two popular yet decisive classification error metrics: True Positive Rate (TPR) and False Positive Rate (FPR).
5. Application of a fuzzy-based ensemble methodology, which is a validated methodology designed to focus on extreme event prediction.

The rest of the manuscript is organised as follows. First, Section 2 describes the fuzzy ensemble methodology proposed in this paper. Then, specific problem definition and databases descriptions are shown in Section 3. Section 4 presents the results obtained for the extreme SWH prediction problem. Finally, some discussion and conclusions about the results are given in Section 5.

2. Proposed methodology

This section details the proposed methodology. First of all, the fuzzy-based cascade ensemble is described as a whole. Then, each specific part of the methodology is presented: Section 2.2 includes a description of the data partition procedure, involving the data partition into fuzzy-soft subsets. Section 2.3 shows the main steps to carry out the prediction. Finally, the evolutionary approach for optimising the proposed fuzzy ensemble is presented in Section 2.4.

2.1. Fuzzy-based cascade ensemble

The concept behind this methodology is to ensemble multiple models of a particular regressor, where the regressor remains the same, but the model fitted varies. Each model is trained with different data, thereby focusing on a specific range of the target variable's value. Therefore, the proposed framework involves decomposing the training data into fuzzy-soft subsets, which are used to train the regressor, and then assemble the prediction of individual models. This sequence is implemented in a multi-layer cascade architecture.

To proceed with, training data is partitioned into fuzzy-soft subsets, this concept was introduced in Molodtsov (1999), and offers a general mathematical tool for dealing with uncertain, fuzzy, not clearly defined objects. A fuzzy subset is described by its Membership Function (MF). This function associates to every sample a real number in the interval $[0, 1]$ known as the pertinence value. This value is interpreted as a degree of a given sample of belonging to a specific fuzzy subset (Cagman et al., 2011). Therefore, each sample is assigned with a pertinence value for each subset, in accordance with the corresponding MF, dependent on the target variable value. Specific details on the fuzzy-soft subsets formation procedure are provided in Section 2.2.

Then, each fuzzy-soft subset is used to train a different regression model, in a way that each model is focused on a specific range of the training data. Afterwards, in order to calculate the predicted value of an incoming test sample, input data is fed to each regressor independently, deriving in a prediction value ($\hat{y}_{\mathcal{M}_i}$) for each model $\mathcal{M}_i$.

These individuals predictions are subsequently introduced into a fuzzy-based ensemble, where the previous layer prediction, together with the MF shapes, are used to determine the pertinence values (p_i) for each prediction. The ensemble output is further weighted with the previous layer's prediction value, on the basis of an hyperparameter named as learning rate (ε). Initially, a regression model ($\mathcal{M}_0$) is trained with all the training data without performing any partition (Layer 0). This model is known as baseline model. Details on the implemented prediction procedure are described in Section 2.3. Being a layered structure, this procedure is repeated sequentially. An example of the framework operation showing the first two layers is provided in Fig. 1.

2.2. Data partition into fuzzy-soft subsets

The developed method employs a cascade architecture that is organised into multiple layers, with the data partition into fuzzy-soft subsets process running within each layer. The parameter m determines the number of groups into which each fuzzy-soft subset from the previous layer (or the total set of training data in the initial layer) is divided. The data partition process is carried out based on the value of the target variable following the procedure below:

1. Data within a given subset are sorted according to the target value.
2. m MFs are defined, each associated with a data subset.
3. The target domain (in percentile) of a specific subset is divided into m regions according to the MFs defined.
4. New data subsets are then formed from each subset based on this domain division.

For the initial layer, the training data is segregated into m groups based on the definition of the MFs. It is noteworthy that these groups may exhibit overlap, as depicted in Fig. 2. In subsequent layers, the same MFs are employed to partition the data subsets from the preceding layer. Consequently, the total number of subgroups belonging to a layer n equals m^n.

The definition of MF shapes plays an important role, making it a critical parameter that influences the success of the algorithm. To address this, a robust evolutionary optimisation algorithm, the Coral Reefs Optimisation Algorithm with Substrate Layers (CRO-SL) (Pérez-Aracil et al., 2023), has been utilised to obtain the most optimal shapes for the MFs (Section 2.4). Five different MF shapes have been considered, involving a bell-shaped, square, exponential (tailed), triangular and double sigmoid function (Eqs. (1), (2), (3), (4), (5)), which can be observed in Fig. 3, where lower and upper limits are represented by q_1 and q_2, respectively.

$$\mathcal{P}_{bell}(i) = \frac{1}{1 + \left|\frac{X_i - c}{a}\right|^{2b}}, \tag{1}$$

where c represents the midpoint of the interval ($c = q_1 + (q_2 - q_1)/2$), a denotes the half of the length between the boundaries ($a = (q_2 - q_1)/2$), and b defines the shape of the function, which has been set to 10.

$$\mathcal{P}_{square}(i) = \begin{cases} 0, & X_i < q_1, \\ 1, & q_1 \le X_i \le q_2, \\ 0, & X_i > q_2. \end{cases} \tag{2}$$

$$\mathcal{P}_{tailed}(i) = \begin{cases} 0.5^{(q_1 - X_i)}, & X_i < q_1, \\ 1, & q_1 \le X_i \le q_2, \\ 0.5^{(X_i - q_2)}, & X_i > q_2. \end{cases} \tag{3}$$

$$\mathcal{P}_{trian}(i) = \begin{cases} 0, & X_i < q_1, \\ \frac{(X_i - q_1)}{(q_2 - q_1)/2}, & q_1 \le X_i \le (q_1 + \frac{(q_2 - q_1)}{2}), \\ \frac{(q_2 - X_i)}{(q_2 - q_1)/2}, & (q_1 + \frac{(q_2 - q_1)}{2}) \le X_i \le q_2, \\ 0, & X_i > q_2, \end{cases} \tag{4}$$

$$\mathcal{P}_{sigmoid}(i) = \begin{cases} \frac{1}{1 + e^{d(-X_i + q_1)}}, & X_i < (q_1 + \frac{(q_2 - q_1)}{2}), \\ \frac{1}{1 + e^{-d(-X_i + q_2)}}, & (q_1 + \frac{(q_2 - q_1)}{2}) \le X_i \end{cases} \tag{5}$$

where d defines the shape of the function and has been set to 20.

The values of parameters b and d have been arbitrarily set for efficiency reasons in the evolutionary tuning process, so that the number of parameters to be tuned is the same regardless of the type of MF.

2.3. Prediction

The presented framework begins with the prediction of an initial regression model ($\mathcal{M}_0$) trained with all the training data without performing any partition (Eq. (6)), which corresponds to the prediction of Layer 0. Note that this model is known as baseline model. Afterwards, the output of a specific layer is computed as the average of the prediction of each model, weighted by its pertinence value, as can be seen in Eqs. (7), (8) and (9) for the first, second, and n layers, respectively. In these equations, i lists all the regression models ($\mathcal{M}_i$) belonging to a layer, $\hat{y}_{\mathcal{M}_i}$ denotes the output of $\mathcal{M}_i$, p_i designates the pertinence value of the evaluated sample for $\mathcal{M}_i$, p_i^* represents the pertinence value of the parent model of $\mathcal{M}_i$ (denoting the model from the previous layer

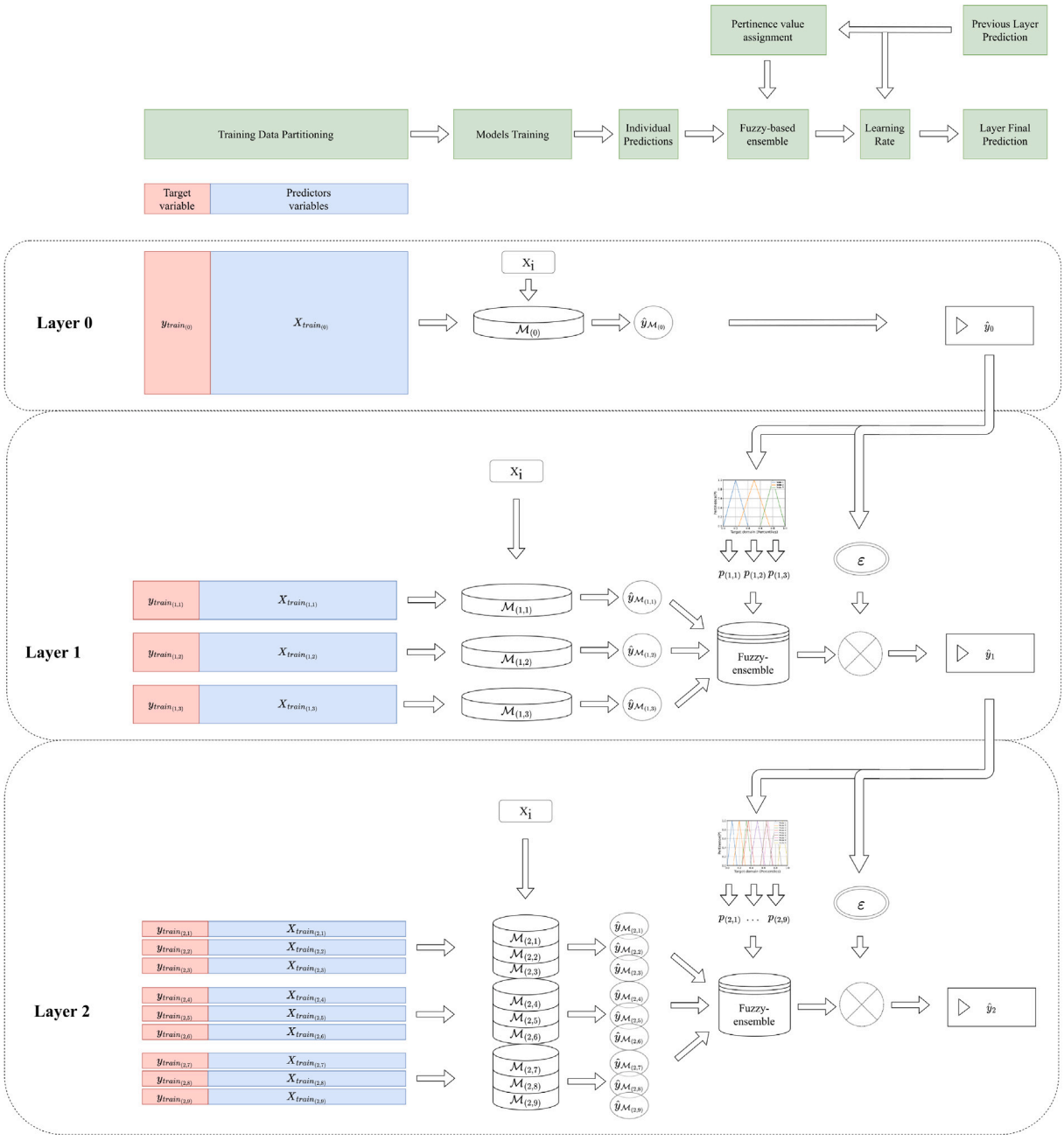
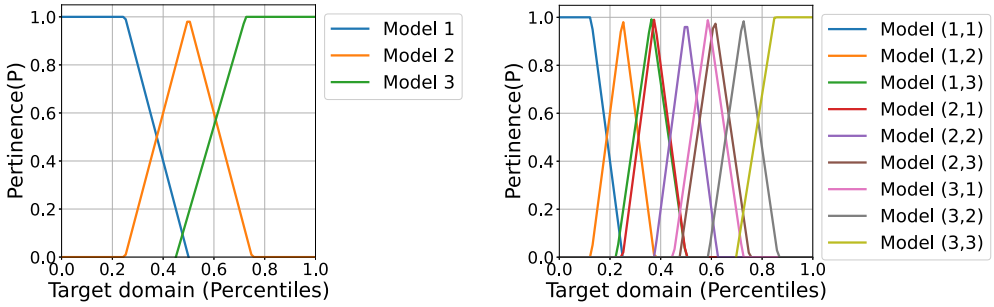


Fig. 1. Fuzzy-based cascade ensemble of regressors architecture.



(a) Layer 1 (b) Layer 2

Fig. 2. Example of triangular MFs for Layers 1 and 2.

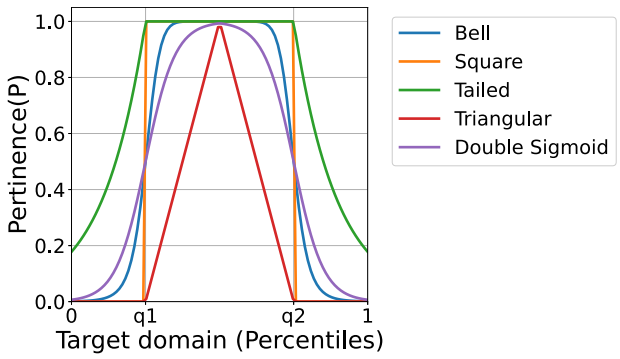


Fig. 3. The five different MF shapes considered.

used to create the corresponding fuzzy-soft subsets) and m represents the number of fuzzy-soft subsets formed on each division.

As shown in Fig. 1, the pertinence values of each layer are obtained according to the previous layer prediction ($\hat{y}_i$) and the MFs associated to each model. For this purpose, the MFs are computed in terms of the target value (instead of the subset percentiles). Fig. 2 illustrates those MFs for the first two layers and the use of the percentile values is shown in Fig. 3 to perform the data partition. Here, entering in those graphics through the x-axis with the value of the previous layer prediction, a pertinence value comprised between 0 and 1 is assigned to each individual model. Also, the pertinence values corresponding to a layer are used in the subsequent layer as p_i^*. Thus, the pertinence value of $\mathcal{M}_1$ in layer 1 (p_1) is used as parent pertinence value of models $\mathcal{M}_{(1,1)}$, $\mathcal{M}_{(1,2)}$ and $\mathcal{M}_{(1,3)}$ in layer 2.

In addition, a learning rate (ε) is defined, aiming at not losing the generalisability of the prediction. Thus, for each layer, the previous layer output is taken into account by a factor of $1-\varepsilon$.

Thus, the output for the first layer is obtained as

$$\hat{y}_0 = \hat{y}_{\mathcal{M}_0}, \tag{6}$$

while the outputs of the subsequent layers (first, second, and nth layer) are given by Eqs. (7), (8), and (9), respectively:

$$\hat{y}_1 = \frac{(\hat{y}_{\mathcal{M}_1}\cdot p_1 + \hat{y}_{\mathcal{M}_2}\cdot p_2 + \hat{y}_{\mathcal{M}_3}\cdot p_3)}{p_1+p_2+p_3}\cdot\varepsilon + \hat{y}_0\cdot(1-\varepsilon) \tag{7}$$

$$\hat{y}_2 = \frac{\sum_{i=1}^{9}(\hat{y}_{\mathcal{M}_i}\cdot p_i\cdot p_i^*)}{\sum_{i=1}^{9} p_i\cdot p_i^*}\cdot\varepsilon + \hat{y}_1\cdot(1-\varepsilon) \tag{8}$$

$$\hat{y}_n = \frac{\sum_{i=1}^{m^n}(\hat{y}_{\mathcal{M}_i}\cdot p_i\cdot p_i^*)}{\sum_{i=1}^{m^n} p_i\cdot p_i^*}\cdot\varepsilon + \hat{y}_{n-1}\cdot(1-\varepsilon) \tag{9}$$

2.4. Hyperparameter tuning via evolutionary optimisation

Given the sensitivity of factors influencing the predicted value, such as the number of MFs (m), their shapes, and the learning rate (ε), evolutionary computation is employed to find an optimal set of values for these parameters. Cases with two, three and four MFs (m) have been tested. In addition, the position and shape of the MFs, together with the learning rate, are optimised via the CRO-SL evolutionary algorithm. The CRO-SL is a multi-method ensemble approach (Wu et al., 2019), based on the CRO algorithm (Salcedo-Sanz et al., 2014). In this multi-method approach, several search operators are applied to a single population, obtaining a powerful evolutionary-based method for optimisation problems. The CRO-SL was initially introduced in Salcedo-Sanz et al. (2016), and the final multi-method ensemble, as the version used in this paper, was introduced in Pérez-Aracil et al. (2022), where a probabilistic-dynamic algorithm was proposed. This version of the CRO-SL is free-access, and a Python code can be obtained via GitHub, as described in Pérez-Aracil et al. (2022).

For each MF, three values are provided, q_1, q_2 and q_3, while the first two are continuous values from 0 to 1 representing the position of the functions according with Eqs. (1), (2), (3), (4), (5), q_3 is a discrete value between 0 and 4 specifying the MF selected for the specific fuzzy-soft subset (Eq. (10)). In order to ensure that the algorithm captures the behaviour of extremes events, the first MF is forced to start at the percentile 0 ($q_1 = 0$) and the last MF is forced to end at the percentile 1 ($q_2 = 1$), in all cases. Therefore, the number of variables to optimise is 5, 8 and 11, for the cases of $m = 2, 3, 4$, respectively.

$$\begin{cases} q_3 = 0 \rightarrow & P_{bell}(i), \\ q_3 = 1 \rightarrow & P_{square}(i), \\ q_3 = 2 \rightarrow & P_{tailed}(i), \\ q_3 = 3 \rightarrow & P_{trian}(i), \\ q_3 = 4 \rightarrow & P_{sigmoid}(i), \end{cases} \tag{10}$$

Since no constraints enforce the overlapping of MFs, there is a possibility that a target variable domain value has an associated pertinence value of 0 for all models. To avoid errors due to division by 0, a term of $0.001\cdot\hat{y_0}$ is added in Eq. (9) when computing the predicted value in layer n. This ensures that even if all membership values are 0 for a specific sample, the prediction for that point will be influenced by the initial model. The updated prediction equation is expressed in Eq. (11).

$$\hat{y}_n = \frac{(\sum_{i=1}^{m^n}(\hat{y}_{\mathcal{M}_i}\cdot p_i\cdot p_i^*)) + 0.001\hat{y_0}}{(\sum_{i=1}^{m^n} p_i\cdot p_i^*) + 0.001}\cdot\varepsilon + \hat{y}_{n-1}\cdot(1-\varepsilon) \tag{11}$$

In order to perform this hyperparameter optimisation, the CRO-SL evolutionary algorithm has been considered, using a cross validation approach on the train data. For this purpose, the full training data (or 70% of the total data) has been divided into 5 validation folds, and the average error encountered in the prediction of these validation data has been used as the fitness function of the optimisation algorithm.

3. Data description and pre-processing

This section describes the data used, which have been gathered from two different sources of information. On the one hand, SWH data has been obtained from the National Data Buoy Center (NDBC) (National Data Buoy Center, 2023). This data is recorded by sensors integrated into marine buoy stations deployed along coastal zones of USA. On the other hand, meteorological and climatological data are obtained from the atmospheric reanalysis project provided by the National Center for Atmospheric Research (NCAR) (Kalnay et al., 1996; Kistler et al., 2001).

As for NDBC data, nine stations deployed in the southwest coast of the USA are considered, whose geographical location is depicted in Fig. 4. Specifically, the SWH data recorded by stations with IDs 46025 and 46026 serve as target variables in two independent cases studies, represented in varying shades of blue in the figure. For each case study, two long-term prediction horizons are taken into account: +24-hour and +36-hour.

Regarding NCAR data, seven reanalysis variables are used as input variables, which are presented in the upper section of Table 1. Given that data from NCAR is available every 2.5° latitude-longitude, for stations with IDs 46025 and 46026, the sub-grid composed of the four closest nodes of reanalysis surrounding their geographical location is used. In this way, the reanalysis variables are computed as the weighted average of the distances from each of the four closest reanalysis nodes to its corresponding station. This computation is done in such a way that the closer the node is to the station, the higher its weighting.

With respect to temporal resolution, both NDBC data and NCAR data corresponding to four years (2017–2020) is selected. Regarding temporal resolution, NDBC provides data every 1 h, whereas NCAR provides data 4 times daily, sampled at 6 h intervals. Therefore, to build the datasets for the two case studies, a data integration process is performed to merge both kind of data. This process is carried out using the SPAMDA software tool (Gómez-Orellana et al., 2021), which

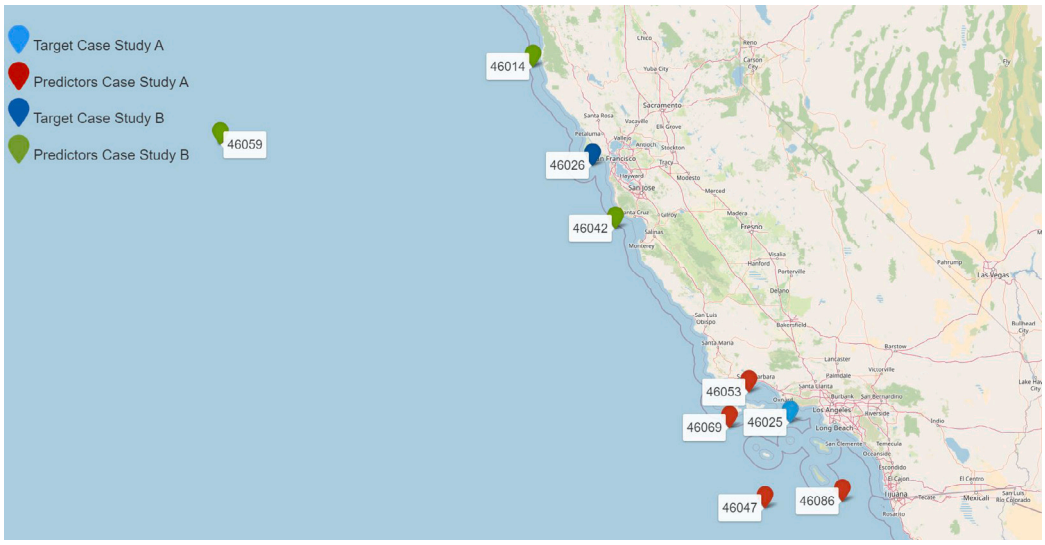


Fig. 4. Geographic locations of the stations considered for target and predictors.

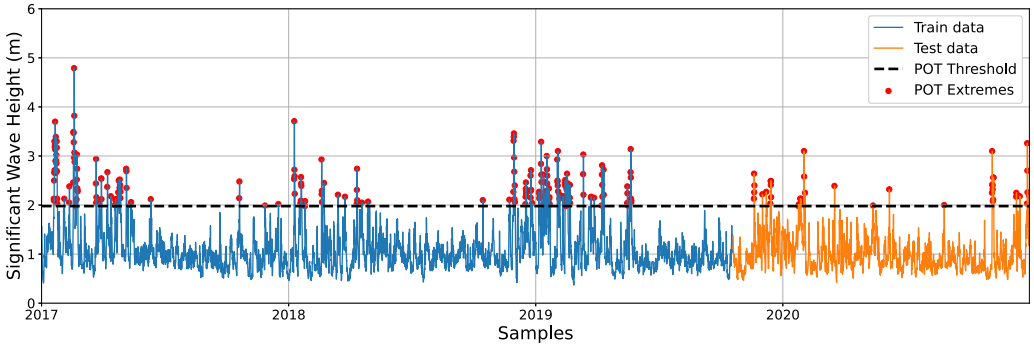


Fig. 5. SWH time series for station 46025.

is designed for this task. Specifically, given that the temporal resolution of reanalysis data is 6 h, the procedure matches reanalysis data with SWH data every 6 h, discarding the remaining SWH data. Hence, the temporal resolution of each instance in the datasets is 6 h. Negligible missing SWH data were filled in using the mean of the five previous and following SWH measurements.

The SWH time series of stations with IDs 46025 and 46026 are depicted in Figs. 5 and 6, respectively. Following the guidelines presented in Prechelt et al. (1994), the dataset of each station is divided as follows: the first 70% of the data, i.e. years 2017–2019, is used to train the proposed methodology and the ML/DL models considered for comparison purposes, while the remaining 25% of data, i.e. year 2020, is used for testing their performance. The Peaks-Over-Threshold (POT) method (Saeed Far and Abd. Wahab, 2016) is considered to define the extremes SWH events. For this purpose, it is first necessary to establish the threshold from which a SWH event is considered extreme, set at the $\text{Mean} + 2 \cdot \text{STD}$ of the SWH values (Viselli et al., 2015), STD being the standard deviation.

In addition to the seven reanalysis variables, SWH data of both nearby stations and the station under study are used to tackle the long-term prediction of SWH. Specifically, for station 46025, prediction involves including the SWH data of four additional stations (red and light blue stations in Fig. 4), whereas for station 46026, prediction involves incorporating the SWH data of five additional nearby stations (green and dark blue stations in Fig. 4), as summarised in the lower section of Table 1. It is important to mention that all the input variables, covering reanalysis data and SWH data, are expressed at +0-h horizon.

A preliminary dataset preparation is performed, consisting on a scaling of the input variables, which is important to ensure that the upper and lower limits of data are in a predefined range. Variable standardisation is performed, causing data to have zero-mean and a unit-variance.

4. Experiments and results

This section describes the details of the experimental work conducted. First, Section 4.1 indicates the error metrics used to assess the performance of the models. Then, Section 4.2 enumerates the ML and DL algorithms used, both within the fuzzy-based ensemble and for assessing the performance of the proposed methodology. Section 4.3 shows the details of the experimental setup considered. Subsequently, Section 4.4 displays the results obtained for the two cases of study. Finally, Section 4.5 includes a discussion and explainability analysis of the results obtained.

4.1. Evaluation metrics

In order to assess the performance of the prediction models, different evaluation metrics have been used, from generic regression error

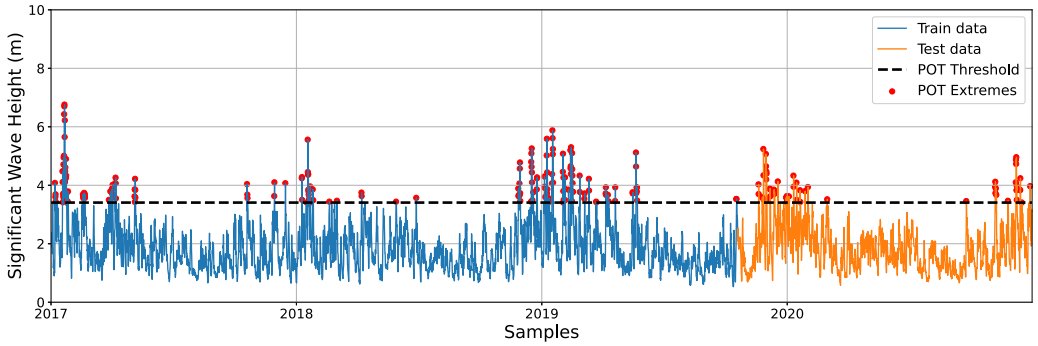


Fig. 6. SWH time series for station 46026.

Table 1
Predictive variables considered for each case study.

Predictive variables used in case studies 46025							
Variable	Description	Minimum	Maximum	Mean	Std deviation	Skewness	Kurtosis
air	Air temperature (K)	277.96	305.96	289.48	4.22	0.32	−0.32
omega	Omega vertical velocity (Pascal/s)	−0.37	0.30	−0.03	0.10	0.03	0.02
pr_wtr	Precipitable water content (kg/m2)	2.30	46.11	16.22	6.67	1.05	1.26
pres	Pressure (Pascal)	95 300.82	98 856.65	97 229.41	355.43	0.14	1.37
rhum	Relative humidity (%)	18.98	100.00	69.76	14.01	−0.27	−0.33
uwnd	Component south-north wind speed (m/s)	−7.88	11.68	2.25	2.55	−0.07	0.37
vwnd	Component west-east wind speed (m/s)	−7.47	12.28	−1.13	1.94	0.56	2.63
SWH25	SWH for station 46025 (m)	0.37	4.79	1.09	0.45	1.94	5.63
SWH47	SWH for station 46047 (m)	0.63	6.62	2.03	0.82	1.41	2.52
SWH53	SWH for station 46053 (m)	0.24	4.59	1.16	0.53	1.47	3.61
SWH69	SWH for station 46069 (m)	0.65	7.68	2.05	0.80	1.45	3.45
SWH86	SWH for station 46086 (m)	0.45	5.24	1.47	0.60	1.78	4.53
Predictive variables used in case studies 46026							
Variable	Description	Minimum	Maximum	Mean	Std deviation	Skewness	Kurtosis
air	Air temperature (K)	278.84	303.50	287.61	3.82	0.52	−0.01
omega	Omega vertical velocity (Pascal/s)	−0.51	0.40	−0.05	0.11	−0.06	0.66
pr_wtr	Precipitable water content (kg/m2)	3.05	46.23	16.72	5.71	0.61	0.69
pres	Pressure (Pascal)	96 693.77	101 411.46	99 617.81	468.25	−0.23	2.31
rhum	Relative humidity (%)	21.86	100.00	76.07	13.03	−0.70	0.24
uwnd	Component south-north wind speed (m/s)	−9.60	10.58	2.29	2.84	−0.54	0.35
vwnd	Component west-east wind speed (m/s)	−11.53	12.84	−1.40	3.10	1.05	1.98
SWH26	SWH for station 46025 (m)	0.53	6.76	1.86	0.79	1.27	2.58
SWH14	SWH for station 46047 (m)	0.58	8.89	2.28	0.95	1.21	2.99
SWH42	SWH for station 46053 (m)	0.58	10.29	2.16	0.93	1.57	4.50
SWH59	SWH for station 46069 (m)	0.76	12.11	2.45	1.17	1.74	5.18

metrics to specific metrics that reflect the models performance in the prediction of extreme SWH events.

First, two common metrics for regression problems, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), have been considered:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{12}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{13}$$

where $\hat{y}$ represents predicted values, y are the observed values and N stands for the number of instances in the dataset. The subscript i is used to refer to a single sample $y_i = y[i]$.

Since the accurate prediction of the extreme SWH events is the scope of this work, the performance of the proposed methodology is evaluated on these extreme SWH data by using two new performance metrics, named Extreme Events Mean Absolute Error (EEMAE) and Extreme Events Root Mean Squared Error (EERMSE) have been used, which correspond to the MAE and RMSE calculation considering only those values.

The main drawback of the EEMAE and EERMSE metrics is that they only focus on the prediction of extreme SWH values, but do not penalise the situation where the model predicts an extreme SWH value, but the observed value is normal (False Positive). These false alarms may lead to severe economic damages, such as a disruption in energy production or the deployment of emergency equipment to reinforce installations when it is not necessary. Two popular classification error metrics are used to account for these cases: True Positive Rate (TPR) and False Positive Rate (FPR).

TPR, also referred to as recall, determines the ability of a model to find all the relevant cases within a dataset. It is computed by dividing the number of relevant cases truly predicted, True Positives (TP), by the total number of relevant cases present in the data, Positives (P). In this context, since we are working with a regression model, we define a threshold above which both actual and predicted values are considered as extreme SWH values (or positive). Consequently, each sample of the test dataset is assigned with a boolean value of TP (1 if both the prediction and the actual value are above T and 0, otherwise, Eq. (14)) and P (1 if the actual value is above T and 0, otherwise, Eq. (15)). Therefore, TPR is computed following Eq. (16), where a value of 1 indicates that all extremes SWH values are predicted correctly and 0 denotes that none have been anticipated.

$$\mathrm{TP}_i = \begin{cases} 0 & \text{if} \quad \hat{y}_i \le T \quad \text{or} \quad y_i \le T \\ 1 & \text{if} \quad \hat{y}_i > T \quad \text{and} \quad y_i > T \end{cases} \tag{14}$$

$$P_i = \begin{cases} 0 & \text{if} \quad \hat{y}_i \leq T \\ 1 & \text{if} \quad \hat{y}_i > T \end{cases} \tag{15}$$

$$\text{TPR} = \frac{\sum_{i=0}^{N} \text{TP}_i}{\sum_{i=0}^{N} P_i} \tag{16}$$

Similarly, the FPR is computed by dividing the number of False Positives (FP), i.e. the number of false alarms or events falsely predicted as extreme SWH, by the number of Negatives (N), i.e. the sum of non-extreme SWH events. According to Eqs. (14) and (15), a boolean value of FP and N is given to each dataset sample (Eqs. (17) and (18), respectively). Then FPR is calculated as shown in Eq. (19), where a value of 0 indicates that all non-extremes SWH events have been predicted correctly, and 1 denotes that all of them have been predicted as false extremes.

$$\text{FP}_i = \begin{cases} 0 & \text{if} \quad \hat{y}_i \leq T \quad \text{or} \quad y_i > T \\ 1 & \text{if} \quad \hat{y}_i > T \quad \text{and} \quad y_i \leq T \end{cases} \tag{17}$$

$$\text{N}_i = \begin{cases} 0 & \text{if} \quad \hat{y}_i > T \\ 1 & \text{if} \quad \hat{y}_i \leq T \end{cases} \tag{18}$$

$$\text{FPR} = \frac{\sum_{i=0}^{N} \text{FP}_i}{\sum_{i=0}^{N} N_i} \tag{19}$$

Finally, G-mean is the root of the product of class-wise sensitivity. This measure tries to maximise the accuracy on each of the classes while keeping these accuracies balanced. G-mean is a good indicators in imbalanced domains because it is independent of the distribution of examples between classes (Barandela et al., 2003). G-mean is computed following the formula shown in Eq. (20), where TPR and TNR (True Negative Rate) are calculated as indicated in Eqs. (16) and (21), respectively, with TN (True Negative) calculated as shown in Eq. (22).

$$\text{G-mean} = \sqrt{\text{TPR} \cdot \text{TNR}} \tag{20}$$

$$\text{TNR} = \frac{\sum_{i=0}^{N} \text{TN}_i}{\sum_{i=0}^{N} N_i} \tag{21}$$

$$\text{TN}_i = \begin{cases} 0 & \text{if} \quad \hat{y}_i > T \quad \text{or} \quad y_i > T \\ 1 & \text{if} \quad \hat{y}_i \leq T \quad \text{and} \quad y_i \leq T \end{cases} \tag{22}$$

4.2. Regression methods

The performance of the fuzzy-based cascade ensemble is assessed by employing different regression algorithms. On the one hand, different regression models have been used as individual models within the fuzzy ensemble. On the other hand, traditional shallow ML methods and state-of-the-art DL methods have been employed for assessing the performance of the proposed methodology. A brief description of these methods is presented here.

4.2.1. Regression methods compared in the fuzzy-based cascade ensemble

Four different ML regression models have been implemented to perform the regression task, looking for models that enable a fast and efficient training and prediction process. These models include two random neural network methods (Extreme Learning Machines (ELM) (Huang et al., 2006), and Random Vector Functional Link network (RVFL) (Shi et al., 2021)), the standard Linear Regression (LR) (Draper and Smith, 1998), and Light Gradient Boosting Machine (LGBM) (Ke et al., 2017). These models are tested independently, i.e. four different instances of the fuzzy-based cascade ensemble are evaluated, each using a different ML methodology for all the individual models within the ensemble, in order to assess the robustness and performance of the methodology.

4.2.2. Algorithms for comparison

A first comparison has been performed with the naive persistence model, meaning that the last known value of wave height is considered as the predicted value.

Ten classical ML methods for regression problems are considered: Regression Trees (RT) (Loh, 2011), Random Forest (RF) (Breiman, 2001), Support Vector Regression (SVR) (Awad and Khanna, 2015), Lasso Regression (Tibshirani, 1996) (Lasso), Least Square Support Vector Regression (LSSVR) (Wang and Hu, 2005), Multilayer Perceptron (MLP) (Gardner and Dorling, 1998), LR, ELM, RVFL and LGBM, these latter four models corresponding to the baseline models from which the proposed fuzzy-based cascade ensemble methodology is developed (models included in layer 0 of the fuzzy-based cascade ensemble).

Among the existing DL methodologies, several approaches are selected for assessing the SWH prediction. These include well-established architectures like Recurrent Neural Networks (RNNs) (Hüsken and Stagge, 2003), Gated Recurrent Units (GRUs) (Chung et al., 2014), Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), and 1-D Convolutional Neural Networks (1D-CNNs). Additionally, we consider hybrid approaches such as the combination of 1D-CNNs and LSTMs (Zhao et al., 2019). Furthermore, we incorporate cutting-edge architectures that have recently gained prominence due to their remarkable performance, such as Residual Networks (ResNet) (Wang et al., 2017) and InceptionTime (Ismail Fawaz et al., 2020).

4.3. Experimental setup

Table 2 shows the parameters used for the benchmark methods considered. In the case of stochastic models, the experiments are repeated for 5 different runs. Furthermore, for the DL methods employed, a dense neural network was implemented after the RNN, GRU, LSTM, 1D-CNN and 1D-CNN + LSTM cases, containing two layers with 64 and 32 neurons in the first three cases and four layers with 256, 128, 64 and 32 neurons in the last two cases. For all this mentioned methods, a batch size of 32 and an early stopping criteria is implemented. Thus, training is stopped when the validation error does not improve in 15 epochs. Moreover, given that DL methods operate on sequential data, the sequence length is treated as a hyperparameter, tailored individually for each database. This tuning process has explored values ranging from 2 to 10.

All of the simulations were run on a Intel(R) Core(TM) i7-10700 CPU with 2.90 GHz and 16 GB RAM using the Python libraries: `sklearn`, `tensorflow`, `scipy` and `tsai`. The stochastic methods have been executed in 5 independent runs to ensure robustness and reliability of results.

4.4. Results

The results achieved when applying the proposed fuzzy-based cascade ensemble methodology are presented below for the two cases studied.

In first place, optimised MFs and learning rates returned by the optimisation algorithm for values of $m = 2$, 3, and 4 are shown in Fig. 7 for case study A (Station 46025) and in Fig. 8 for case study B (Station 46026), considering in both cases a prediction horizon of +24-h. It can be observed that different strategies are selected for each case, exhibiting differing degrees of overlap. It should be reminded that the first fuzzy-soft subset is fixed to start at percentile 0, while the last subset is fixed to end at percentile 1. This setup ensures the inclusion of extreme SWH events within the analysis. In both cases, it is observed that the higher learning rates are obtained by the LR models.

Subsequently, for the three values of MF considered ($m = 2$, $m = 3$ and $m = 4$), the optimal number of layers of the cascade ensemble (n) is determined according to the performance on the training data. Fig. 9 shows the evolution of the training error metrics when increasing the

Table 2
Experimental setup.

ML Methods	**RT**		**RF**	
	max depth	400	n estimators	400
	SVR		**LSSVR**	
	regularisation parameter	400	gamma	0.01
	epsilon	0.1	n estimators	400
	MLP		**ELM**	
	hidden layers	2	hidden size	500
	neurons per layer	64		
	activation	'relu'		
	solver	'adam'		
	RVFL		**LGBM**	
	hidden nodes	2000	num leaves	31
	regularisation parameter	0.001	n estimators	100
DL Methods	**RNN**		**GRU**	
	number of layers	2	number of layers	1
	neurons per layer	64	neurons per layer	64
	LSTM		**1D-CNN**	
	number of layers	2	number of CNN layers	1
	neurons per layer	64,32	filters	128
			kernel size	4
	ResNet		**InceptionTime**	
	batch size	512	batch size	512
	epochs	15	epochs	15

number of layers for the different values of m and the LR model for the first case study. A clear trend is observed in this figure: the metric related to extreme SWH values (EEMAE) substantially decreases in the initial layers at the expense of a slight deterioration of non-extremes SWH values related metric (MAE). This tendency is equally observed in the remaining ML models and in both cases of study. Therefore, the optimal number of layers is selected based on the sum of the MAE and EEMAE metrics, searching for the most balanced performance.

Then, results for the specific models and values of m are shown in Tables 3 and 4 for both prediction horizons considered in each case of study. It can be noted that the application of the fuzzy ensemble methodology yields consistent trends across different horizons and models: (1) a notable improvement in terms of prediction metrics for extreme values (EEMAE, EERMSE and TPR) is observed compared to the baseline models, representing the standard ML model trained with the entire dataset (Layer 0); (2) this is accompanied by a slight deterioration in the standard metrics (MAE and RMSE) and in the FPR. This outcome is inherent to addressing problems associated with predicting extreme events (Torgo et al., 2013; Peláez-Rodríguez et al., 2022); (3) however, the significant improvement relies in the enhancement observed in the metrics related to the overall model performance, achieving a balance between extreme and non-extreme SWH predictions (Sum of MAEs and G-Mean). Also, regarding the number of MFs selected for each model (m), no definitive conclusions can be drawn as a similar performance is observed across all cases. Table 5 shows the average sum of MAEs metric, which is precisely the one used in the optimisation process, for the four ML models across all the cases under study. It can be observed how in the first station, models perform better when increasing the number of MFs. While for the second case, it can be concluded that the impact of this parameter is relatively minor, with changes below 4% between best and worst case.

Finally, Fig. 10 show the MAE and EEMAE error metrics for the ML and DL methods used as benchmarks in comparison with the fuzzy-based ensemble models. It is possible to observe the existing trade-off between a strong prediction on the extremes SWH events (low EEMAE) and a robust overall prediction (MAE). For instance, it can be noted how the DL models, which are the best in the MAE metric, obtain the worst results for the extreme SWH events. In this regard, the proposed methodology exhibits the most balanced performance in both metrics, with a slight worsening of the MAE compared to the base ML models, but with a significant improvement in the prediction of extreme SWH values. It is worth mentioning that the optimisation of the methodology parameters (number of layers, learning rate and shape of the MFs) is conducted with the objective of minimising the sum of these two metrics (MAE and EEMAE), being precisely in this combined metric where these models exploit their full potential.

The observed trade-off between the metrics, with a slight decrease in MAE performance when improving EEMAE, is expected when tackling extreme events prediction. While the use of increasingly specialised individual models within the ensemble learning framework enhances performance in that specific range of the data domain, this comes at the cost of a slight deterioration in overall performance.

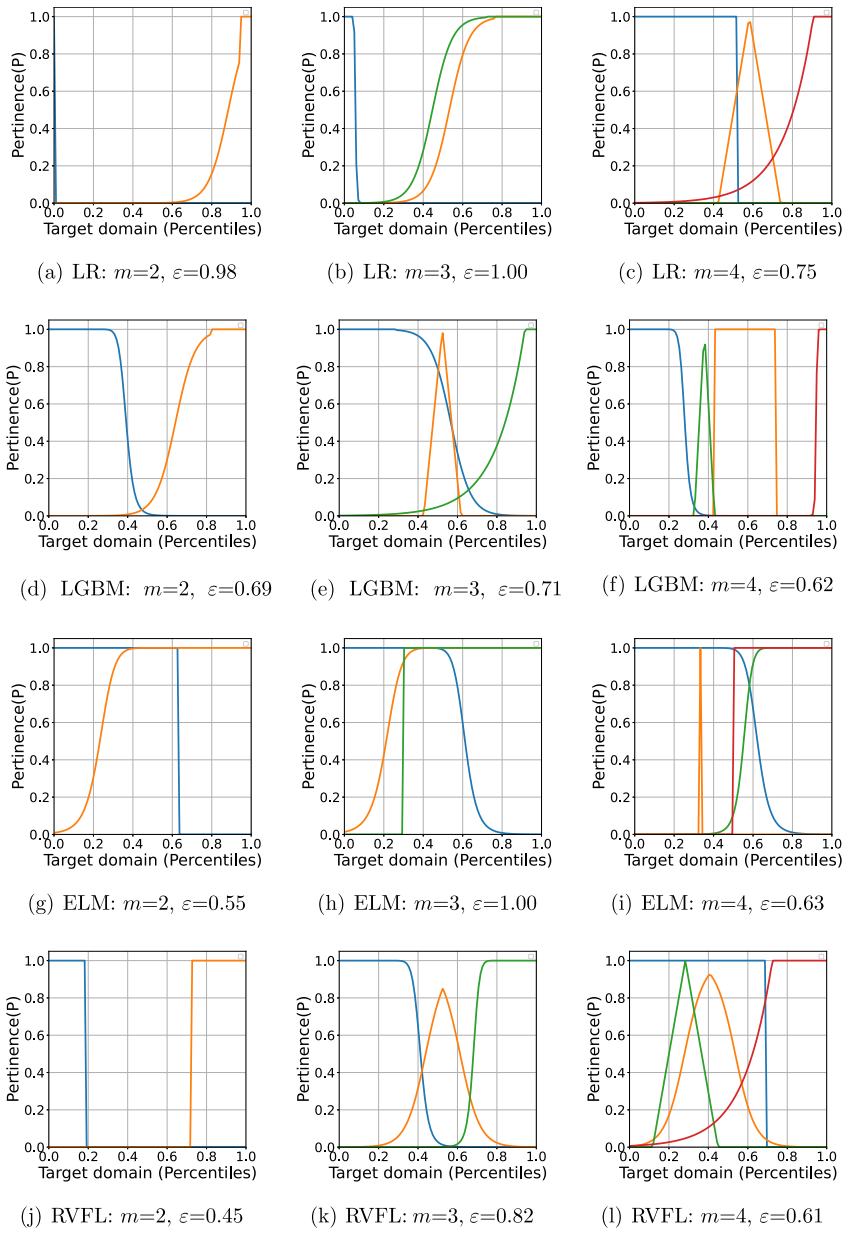
4.5. Explainability analysis and discussion

One of the main advantages of the proposed fuzzy ensemble methodology lies in its transparency, which allows an exhaustive analysis of the explainability of the method in order to fully understand the way in which the prediction process is being performed. In this section, the explainability of the method is addressed for both cases under study.

4.5.1. Case study A: Station 46025

One specific case out of the 12 assessed, comprising different models and values of m, is chosen for a comprehensive analysis of its behaviour. Given that the tuning of the MF is conducted through evolutionary computation minimising the sum of the MAE and the EEMAE, the model exhibiting the best performance in this combined metric is selected. Specifically, the explainability analysis is performed on the LGBM model using 4 MFs, which achieves a value of 0.9 in this MAE+EEMAE metric for the 24-hour prediction horizon.

The analysis begins by examining all the models involved in the prediction process using a fuzzy ensemble. The model under study adopts a total of 2 layers ($n = 2$), where each layer subsequently divides the training data subsets into four groups based on the forms of the four MFs ($m = 4$) defined by the optimisation algorithm. Subsequently, each of these subsets is utilised to train a specific model within the ensemble, from this point on, these individual models are referred to in sequential order starting with 0. Therefore, the initial layer consists of one model trained with all the available training data (Model 0). Then, layer one divides this data into four groups according to the MFs shapes (Models 1–4). Finally, in the second layer, the training data is divided into 16



(a) LR: m=2, ε=0.98 (b) LR: m=3, ε=1.00 (c) LR: m=4, ε=0.75

(d) LGBM: m=2, ε=0.69 (e) LGBM: m=3, ε=0.71 (f) LGBM: m=4, ε=0.62

(g) ELM: m=2, ε=0.55 (h) ELM: m=3, ε=1.00 (i) ELM: m=4, ε=0.63

(j) RVFL: m=2, ε=0.45 (k) RVFL: m=3, ε=0.82 (l) RVFL: m=4, ε=0.61

Fig. 7. Optimised MFs for case study A and different values of m for a +24-h prediction horizon Each line represents a MF returned by the optimisation algorithm, with no specific order.

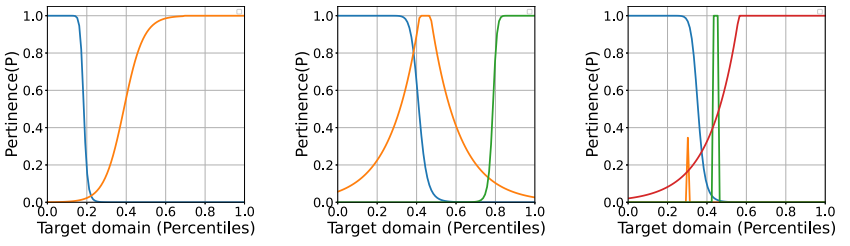
data subsets (Models 5–21). The data partitioning process is depicted in Figs. 11 and 12. The former represents the shapes of the MFs considered for each model, while the latter displays the limits used in the data partition process ($q1$ and $q2$ in Fig. 3). It is important to note that this partitioning of the training dataset is based on the values of the target variable, with the aim of training specific models for each range within the target domain.

Then, following Eq. (11) it is possible to calculate the contribution of each model to the predicted value, considering the pertinence values of each model (p) and the learning rate (ε), which for the case under study is equal to 0.62 (Fig. 7(f)). This contribution may be computed for each sample. Table 6 shows the average contribution of each model considering both the entire test data and only the extreme SWH events within it. Additionally, the table displays the limits used to partition the training dataset to obtain specific data subsets for each model.

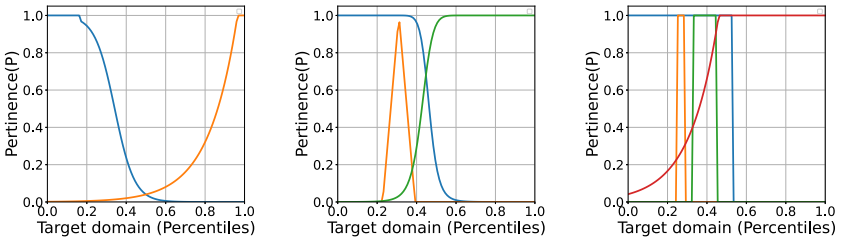
It may be observed that certain models exhibit greater importance than others in the prediction task. Moreover, the significance of these models varies notably when predicting extreme events. In this way, it is possible to examine the distinctive attributes of the models identified as most significant. The five models with higher contribution are selected for further examination in Section 4.5.3. Figs. 13 and 14 showcase the corresponding MFs associated with these models, along with the normalised averages of each predictor in the specific training data subset of each model.

4.5.2. Case study B: Station 46026

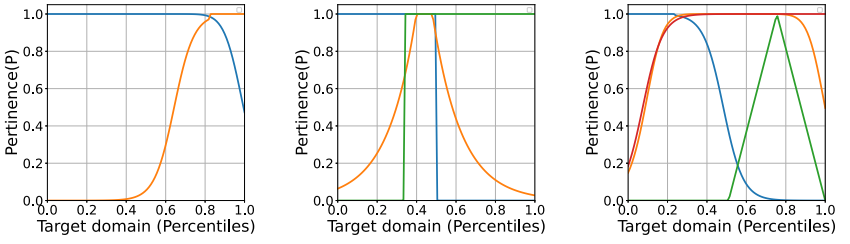
Regarding the explainability of the ensemble, the case with the lowest error in terms of the MAE+EEMAE metric has been selected to conduct this analysis, corresponding to the LGBM regression model, with a configuration of 2 MFs ($m = 2$) and 3 layers ($n = 3$), meaning that a total of 15 models are involved in the prediction process (1+2+4+8), for the purpose of clarity in the following figures, these models are labelled sequentially, starting from 0. The MFs of this individual models are shown in Fig. 15, and the corresponding portions of data used for the training of the models are displayed in Fig. 16, where it is possible to see how the models become more specific when increasing the depth of the ensemble.



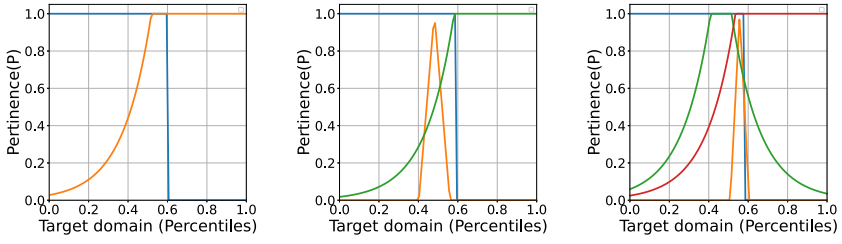
(a) LR: m=2, ε=0.99 (b) LR: m=3, ε=0.95 (c) LR: m=4, ε=1.00



(d) LGBM: m=2, ε=0.66 (e) LGBM: m=3, ε=0.70 (f) LGBM: m=4, ε=0.55

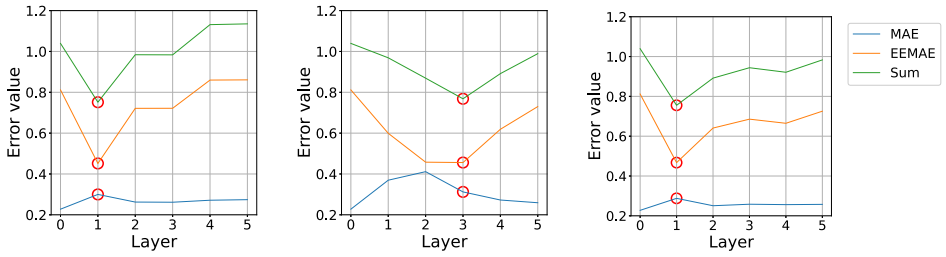


(g) ELM: m=2, ε=0.46 (h) ELM: m=3, ε=0.45 (i) ELM: m=4, ε=0.89



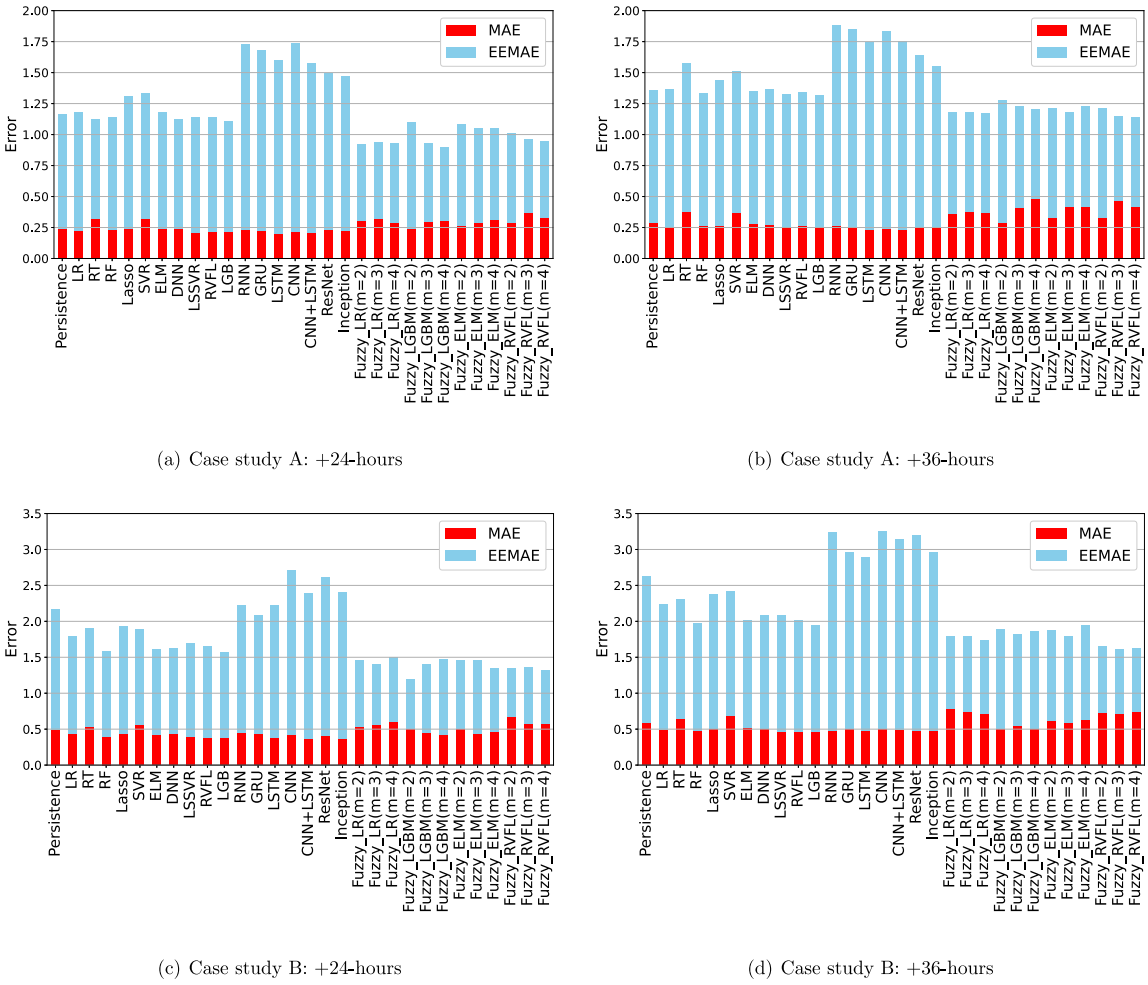
(j) RVFL: m=2, ε=0.51 (k) RVFL: m=3, ε=0.82 (l) RVFL: m=4, ε=0.82

Fig. 8. Optimised MFs for case study B and different values of m. Each line represent a MF returned by the optimisation algorithm, with no specific order.



(a) LR: $m = 2$ (b) LR: $m = 3$ (c) LR: $m = 4$

Fig. 9. Error metrics evolution on training data when increasing the number of layers for LR and $m = 2$, $m = 3$ and $m = 4$ in case study A for a +24-h prediction horizon.



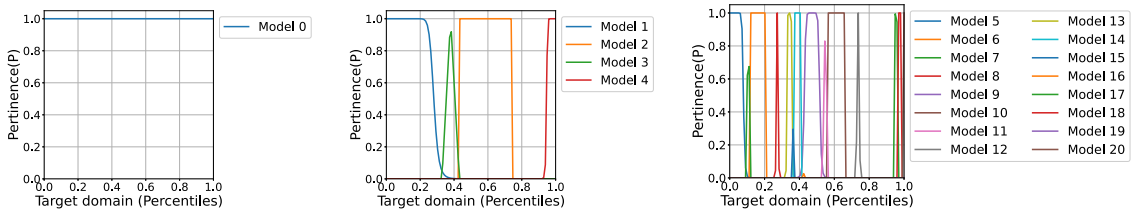
(a) Case study A: +24-hours

(b) Case study A: +36-hours

(c) Case study B: +24-hours

(d) Case study B: +36-hours

Fig. 10. MAE and EEMAE error metrics for the benchmarks algorithms.

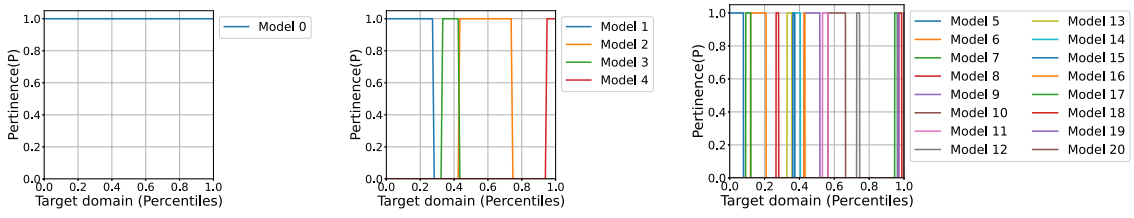


(a) Layer 0

(b) Layer 1

(c) Layer 2

Fig. 11. MFs shapes for the different models involved in each layer.



(a) Layer 0

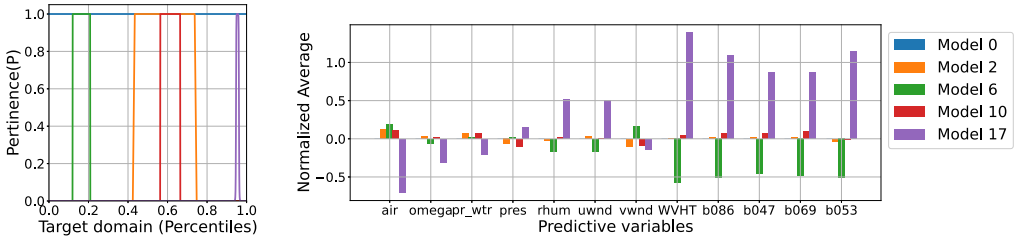
(b) Layer 1

(c) Layer 2

Fig. 12. Training data partition made for the different models involved in each layer.

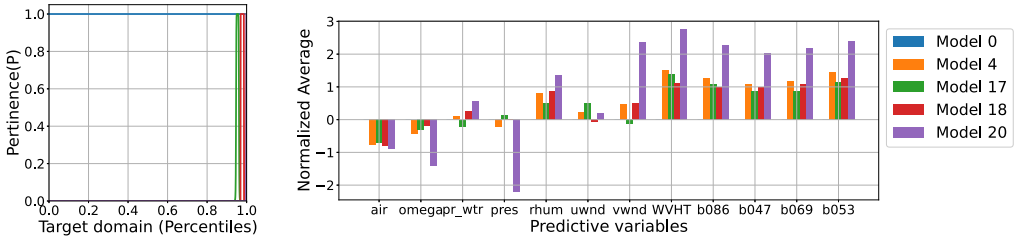
Table 3
Error metrics for the proposed methodology applied on case study A.

+24-h		MAE	EEMAE	Sum	RMSE	EERMSE	TPR	FPR	G-Mean
LR	Baseline model	**0.22**	0.96	1.18	**0.30**	1.06	0.05	**0.00**	0.08
	2 MFs ($n = 1$)	0.30	0.62	0.92	0.44	0.85	0.41	0.09	0.16
	3 MFs ($n = 3$)	0.32	0.62	0.94	0.44	0.83	0.39	0.08	0.16
	4 MFs ($n = 1$)	0.29	0.64	0.93	0.43	0.88	0.41	0.09	0.16
LGBM	Baseline model	**0.22**	0.89	1.11	0.31	1.02	0.15	0.01	0.19
	2 MFs ($n = 5$)	0.24	0.86	1.10	0.33	0.98	0.17	0.02	0.18
	3 MFs ($n = 3$)	0.30	0.63	0.93	0.42	0.85	**0.56**	0.13	0.16
	4 MFs ($n = 2$)	0.31	**0.59**	**0.90**	0.45	0.84	**0.56**	0.15	0.15
ELM	Baseline model	0.24	0.94	1.18	0.33	1.06	0.07	0.01	0.11
	2 MFs ($n = 4$)	0.26	0.82	1.08	0.37	0.99	0.17	0.03	0.13
	3 MFs ($n = 3$)	0.29	0.76	1.05	0.39	0.90	0.20	0.02	0.18
	4 MFs ($n = 2$)	0.31	0.74	1.05	0.42	0.88	0.17	0.04	0.11
RVFL	Baseline model	**0.22**	0.92	1.14	**0.30**	1.03	0.07	**0.00**	0.12
	2 MFs ($n = 2$)	0.29	0.72	1.01	0.41	0.94	0.37	0.04	**0.23**
	3 MFs ($n = 2$)	0.37	**0.59**	0.96	0.49	**0.79**	0.46	0.12	0.14
	4 MFs ($n = 2$)	0.33	0.62	0.95	0.48	0.88	0.54	0.15	0.14
+36-h		**MAE**	**EEMAE**	**Sum**	**RMSE**	**EERMSE**	**TPR**	**FPR**	**G-Mean**
LR	Baseline model	**0.25**	1.12	1.37	**0.34**	1.19	0.00	**0.00**	0.00
	2 MFs ($n = 2$)	0.36	0.82	1.18	0.49	0.97	0.27	0.09	0.11
	3 MFs ($n = 3$)	0.38	0.80	1.18	0.49	0.94	0.24	0.10	0.09
	4 MFs ($n = 1$)	0.37	0.80	1.17	0.54	1.00	0.39	0.14	0.11
LGBM	Baseline model	0.26	1.06	1.32	0.35	1.15	0.05	0.01	0.07
	2 MFs ($n = 5$)	0.29	0.99	1.28	0.38	1.09	0.07	0.01	0.10
	3 MFs ($n = 1$)	0.41	0.82	1.23	0.56	1.03	0.39	0.13	0.11
	4 MFs ($n = 3$)	0.48	0.72	1.20	0.63	0.94	**0.56**	0.28	0.09
ELM	Baseline model	0.28	1.07	1.35	0.37	1.17	0.00	**0.00**	0.00
	2 MFs ($n = 4$)	0.33	0.88	1.21	0.46	1.03	0.24	0.06	0.13
	3 MFs ($n = 4$)	0.42	0.76	1.18	0.52	0.90	0.29	0.07	**0.14**
	4 MFs ($n = 4$)	0.42	0.81	1.23	0.52	0.94	0.17	0.06	0.09
RVFL	Baseline model	0.26	1.08	1.34	**0.34**	1.16	0.00	**0.00**	0.00
	2 MFs ($n = 2$)	0.33	0.88	1.21	0.47	1.05	0.27	0.08	0.12
	3 MFs ($n = 3$)	0.47	**0.68**	1.15	0.62	**0.87**	0.34	0.16	0.09
	4 MFs ($n = 3$)	0.42	0.72	**1.14**	0.57	0.91	0.32	0.15	0.09



(a) MFs of each model (b) Normalised average of predictors for each model training data

Fig. 13. Analysis of most important models considering entire test data.

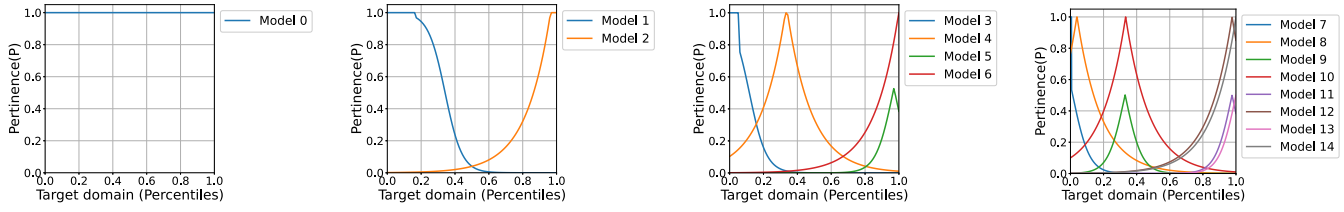


(a) MFs of each model (b) Normalised average of predictors for each model training data

Fig. 14. Analysis of most important models considering only extreme events within test data.

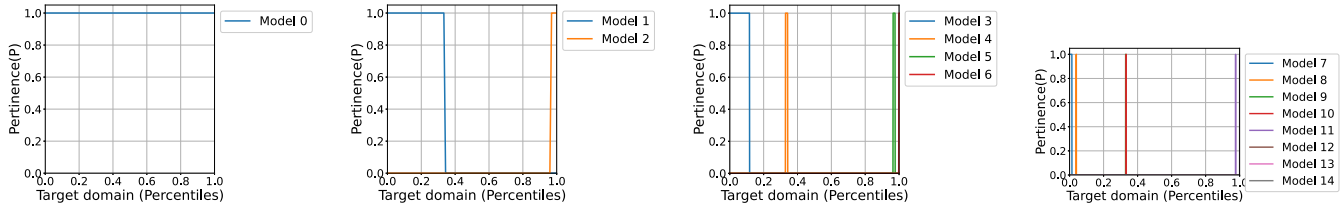
Table 4
Error metrics for the proposed methodology applied on case study B.

+24-h		MAE	EEMAE	Sum	RMSE	EERMSE	TPR	FPR	G-Mean
LR	Baseline model	0.43	1.36	1.79	0.56	1.59	0.13	**0.01**	0.19
	2 MFs ($n = 3$)	0.53	0.92	1.45	0.69	1.29	0.52	0.07	0.29
	3 MFs ($n = 1$)	0.56	0.84	1.40	0.73	1.20	0.50	0.06	0.28
	4 MFs ($n = 1$)	0.60	0.90	1.50	0.76	1.18	0.31	0.02	**0.34**
LGBM	Baseline model	**0.38**	1.19	1.57	0.52	1.40	0.15	**0.01**	0.22
	2 MFs ($n = 3$)	0.50	0.69	**1.19**	0.68	**1.05**	**0.72**	0.13	0.25
	3 MFs ($n = 4$)	0.44	0.96	1.40	0.60	1.22	0.41	0.05	0.28
	4 MFs ($n = 1$)	0.42	1.05	1.47	0.56	1.28	0.17	**0.01**	0.22
ELM	Baseline model	0.42	1.19	1.61	0.57	1.39	0.19	**0.01**	0.23
	2 MFs ($n = 4$)	0.50	0.96	1.46	0.66	1.15	0.31	0.03	0.26
	3 MFs ($n = 2$)	0.44	1.02	1.46	0.58	1.23	0.35	0.02	0.33
	4 MFs ($n = 4$)	0.47	0.88	1.35	0.62	1.07	0.41	0.04	0.32
RVFL	Baseline model	**0.38**	1.28	1.66	**0.51**	1.51	0.11	**0.01**	0.16
	2 MFs ($n = 4$)	0.67	**0.67**	1.34	0.89	1.06	0.63	0.16	0.19
	3 MFs ($n = 2$)	0.58	0.78	1.36	0.77	1.09	0.37	0.05	0.26
	4 MFs ($n = 3$)	0.57	0.75	1.32	0.76	1.11	0.67	0.13	0.23
+36-h		**MAE**	**EEMAE**	**Sum**	**RMSE**	**EERMSE**	**TPR**	**FPR**	**G-Mean**
LR	Baseline model	0.49	1.75	2.24	0.64	1.87	0.00	**0.00**	0.00
	2 MFs ($n = 1$)	0.78	1.01	1.79	0.99	1.35	0.17	0.10	0.08
	3 MFs ($n = 1$)	0.74	1.05	1.79	0.92	1.35	0.11	0.07	0.07
	4 MFs ($n = 1$)	0.72	1.02	1.74	0.92	1.36	0.23	0.10	0.10
LGBM	Baseline model	**0.47**	1.48	1.95	0.63	1.67	0.02	**0.00**	0.03
	2 MFs ($n = 7$)	0.51	1.38	1.89	0.66	1.57	0.06	0.01	0.09
	3 MFs ($n = 3$)	0.54	1.28	1.82	0.69	1.50	0.06	0.01	0.08
	4 MFs ($n = 4$)	0.50	1.36	1.86	0.65	1.57	0.04	0.01	0.06
ELM	Baseline model	0.52	1.49	2.01	0.69	1.67	0.08	0.01	0.11
	2 MFs ($n = 4$)	0.61	1.26	1.87	0.81	1.55	0.26	0.04	**0.21**
	3 MFs ($n = 3$)	0.59	1.20	1.79	0.77	1.36	0.15	0.03	0.14
	4 MFs ($n = 3$)	0.63	1.32	1.95	0.81	1.53	0.17	0.03	0.16
RVFL	Baseline model	**0.47**	1.55	2.02	**0.62**	1.69	0.02	**0.00**	0.03
	2 MFs ($n = 2$)	0.73	0.93	1.66	0.95	1.32	0.43	0.16	0.13
	3 MFs ($n = 3$)	0.72	0.90	**1.62**	0.95	1.34	**0.62**	0.23	0.14
	4 MFs ($n = 3$)	0.74	0.89	1.63	**0.89**	**1.29**	0.53	0.17	0.15



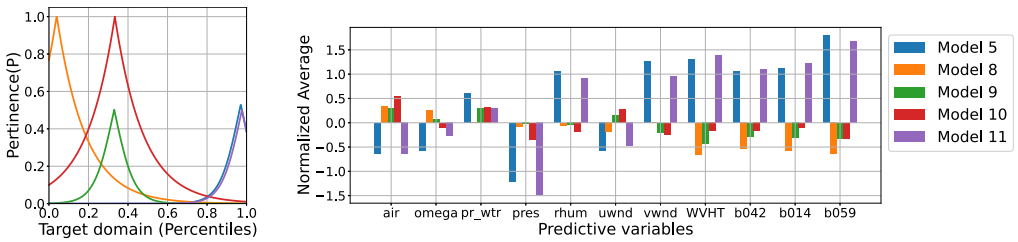
(a) Layer 0 (b) Layer 1 (c) Layer 2 (d) Layer 3

Fig. 15. MF shapes for the different models involved in each layer.



(a) Layer 0 (b) Layer 1 (c) Layer 2 (d) Layer 3

Fig. 16. Training data partition made for the different models involved in each layer.



(a) MFs of each model (b) Normalised average of predictors for each model training data

Fig. 17. Analysis of most important models considering entire test data.

Table 5
Average Sum of MAEs.

	2 MFs	3 MFs	4 MFs
Station 46025 (+24-h)	1.028	0.970	**0.958**
Station 46025 (+36-h)	1.220	**1.185**	**1.185**
Station 46026 (+24-h)	**1.360**	1.405	1.410
Station 46026 (+36-h)	1.803	**1.755**	1.795

Table 6
Average contribution of each model considering the total test data and only extreme events, for the LGBM model with $m = 4$ and $n = 2$.

Layer	Model	Inferior Limit (m)	Superior Limit (m)	Average contribution	Average contribution (extremes)
0	0	0.37	4.79	**0.145**	**0.145**
1	1	0.37	0.82	0.068	0.014
	2	0.92	1.22	**0.089**	0.045
	3	0.85	0.92	0.027	0.009
	4	1.96	4.79	0.053	**0.167**
2	5	0.37	0.66	0.048	0.000
	6	0.70	0.77	**0.086**	0.030
	7	0.68	0.70	0.013	0.000
	8	0.81	0.82	0.019	0.010
	9	0.92	0.98	0.020	0.007
	10	1.02	1.11	**0.139**	0.045
	11	0.99	1.02	0.039	0.035
	12	1.20	1.22	0.000	0.000
	13	0.85	0.87	0.024	0.012
	14	0.88	0.91	0.037	0.000
	15	0.87	0.88	0.010	0.012
	16	0.92	0.92	0.000	0.000
	17	1.96	2.10	**0.147**	**0.267**
	18	2.18	2.60	0.026	**0.121**
	19	2.13	2.18	0.005	0.006
	20	3.30	4.79	0.007	**0.076**

Next, Table 7 shows the average contribution of each model for predicting the test samples along with the average contribution specifically for predicting extreme events. Finally, the models with higher importance in both cases are selected for further analysis (Figs. 17 and 18) in Section 4.5.3.

4.5.3. Discussion

Once the explainability analysis has been performed for both cases of study, it is possible to analyse the results and extract some common conclusions.

These analyses highlight the differences between the selected models for the entire test dataset (Figs. 13 and 17) and only for extreme events (Figs. 14 and 18): (1) While the data subsets of the most important models for the overall test dataset appear to be distributed across

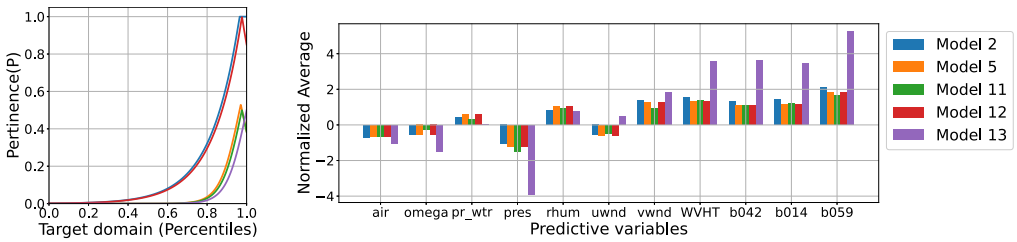
Table 7
Average contribution of each model considering the total test data and only extreme events, for the LGBM model with $m = 2$ and $n = 3$.

Layer	Model	Inferior Limit (m)	Superior Limit (m)	Average contribution	Average contribution (extremes)
0	0	0.53	6.76	0.039	0.039
1	1	0.53	1.42	0.045	0.005
	2	3.50	6.76	0.031	**0.071**
2	3	0.53	1.05	0.073	0.003
	4	1.40	1.42	0.083	0.008
	5	3.50	3.76	**0.102**	**0.199**
	6	5.65	6.76	0.002	0.015
3	7	0.53	0.87	0.022	0.002
	8	1.05	1.05	**0.106**	0.007
	9	1.40	1.41	**0.139**	0.011
	10	1.42	1.42	**0.106**	0.007
	11	3.50	3.60	**0.269**	**0.329**
	12	3.50	3.76	0.043	**0.253**
	13	5.65	6.76	0.007	**0.052**
	14	5.65	6.76	0.000	0.000

the entire target variable domain (Figs. 13(a) and 17(a)), the analysis of extreme events reveals that the models with higher relevance are precisely those trained with extreme events (Figs. 14(a) and 18(a)), along with the baseline model trained with the entire training database (Model 0), which appears as one of the five most important in the first case of study. (2) Then, regarding the distribution of predictive data in the training data subsets of each model, it is observed that the models specialised in extreme events are trained with specific data that differs from the average in some predictive variables: air temperature, omega and pressure are lower than the average in these subsets, meaning that an unusually low value may indicate the presence of extreme waves in the following hours. Also, an abnormally high value of humidity, wind speed and nearby SWH in preceding hours may drive to extreme wave height events. The conclusions obtained are aligned with the findings present in the state-of-the-art (Demetriou et al., 2021). Furthermore, to the authors' knowledge, this is the first time that an analysis of this type is performed distinguishing between the overall events and only the extreme events, seeking to find those predictor variables that have a higher significance in these later cases.

5. Conclusions

In the field of offshore renewable energy and wave energy harvesting, providing precise predictions of the SWH represents a crucial factor for effective sector management. In this study, our focus is on predicting extreme SWH, which holds a critical role in coastal engineering endeavours and carries significant geophysical implications. With the aim of overcoming the inherent problem of skewness and



(a) MFs of each model (b) Normalised average of predictors for each model training data

Fig. 18. Analysis of most important models considering only extreme events within test data.

imbalance associated with the prediction of these extreme SWH events, a fuzzy-based cascade ensemble of regression models is proposed. This methodology allows to remarkably improve the predictive performance of extreme SWH events, by using different models specialised in different ranges of the target domain. Two case studies have been considered, conducting the prediction on two stations located at the southwest coast of the USA. Reanalysis data variables providing information on various meteorological factors are used along with SWH measurements obtained from the nearby stations and the station under examination. Two long-term prediction horizons are considered (+24-h and +36-h).

A striking compromise has been observed in all the assessed cases, balancing between increasing the performance in predicting extreme SWH events and maintaining a correct overall prediction. In this sense, the results provided by the fuzzy ensemble methodology outperform all other ML and DL models considered as benchmarks, achieving high rates of extreme SWH detection while maintaining the number of false positives low.

In addition, one of the novelties of the proposed approach lies in the no requirement of any data balancing technique to solve the challenge of extreme events forecasting. Instead, it is based on the ensemble of different regression models, where the key is to correctly divide the training data in subsets so that each model is focused on a specific part of the data spectrum. This optimal data partitioning is specific to the problem: the shape, position and overlapping of the MFs depend on the database, as well as on the degree of imbalanceness of the problem. The determination of these parameters of the methodology is performed by means of a powerful, well-established state-of-the-art optimisation algorithm (CRO-SL).

Furthermore, the simplicity and transparency provided by this fuzzy-based methodology enable the extraction of explainable insights regarding the significant contributions of the specific fuzzy sets to the prediction process. This facilitates an understanding of how these models differ from others in terms of the distribution of the predictive data used to train each model. Examining the models that contribute most to predicting extreme SWH reveals the influential predictors driving accurate forecasts of these samples. These include humidity, wind speed, and SWH from nearby buoys in preceding hours. Thus, higher values than the average of these variables may indicate an increased potential risk of the occurrence of extreme SWH in the following hours. Conversely, unusually low air temperatures and pressures can also serve as crucial indicators for the occurrence of extreme events.

Future research lines expanding this work include improving the fuzzy-based ensemble methodology by applying a multi-objective optimisation algorithm, aiming at finding the pareto front between the models that perform best in extremes SWH events, thus achieving the minimum EEMAE, and those that perform best in non-extreme SWH values, which are measured by reaching the minimum MAE.

CRediT authorship contribution statement

C. Peláez-Rodríguez: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Conceptualization. **J. Pérez-Aracil:** Writing – review & editing, Validation, Supervision, Investigation. **A.M. Gómez-Orellana:** Writing – review & editing, Validation, Supervision. **D. Guijo-Rubio:** Writing – review & editing, Validation, Supervision. **V.M. Vargas:** Writing – review & editing, Validation, Supervision. **P.A. Gutiérrez:** Supervision, Project administration, Funding acquisition. **C. Hervás-Martínez:** Supervision, Project administration, Investigation. **S. Salcedo-Sanz:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The present study has been supported by the European Commission, projects “Test and Experiment Facilities for the Agri-Food Domain, AgriFoodTEF” (grant ref.: DIGITAL-2022-CLOUD-AI-02, 101100622) and “CLImate INTelligence: Extreme events detection, attribution and adaptation design using machine learning, CLINT” (grant ref.: H2020-LC-CLA-2020-2, 101003876), by the ENIA International Chair in Agriculture, University of Córdoba (grant ref.: TSI-100921-2023-3), by the “Agencia Estatal de Investigación (España)”, Spanish Ministry of Research and Innovation (grant refs.: PID2023-150663NB-C21 and PID2023-150663NB-C22/AEI/10.13039/501100011033), and by the University of Córdoba and Junta de Andalucía (grant ref.: PP2F_L1_15). Antonio Manuel Gómez-Orellana has been supported by “Consejería de Transformación Económica, Industria, Conocimiento y Universidades de la Junta de Andalucía” (grant ref.: PREDOC-00489). David Guijo-Rubio has been supported by the “Agencia Estatal de Investigación (España)” MCIU/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR (grant ref.: JDC2022-048378-I).

Data availability

The datasets used in this work are publicly available.

References

Abbas, M., Min, Z., Liu, Z., Zhang, D., 2024. Unravelling oceanic wave patterns: A comparative study of machine learning approaches for predicting significant wave height. Appl. Ocean Res. 145, 103919.

Afzal, M.S., Kumar, L., Chugh, V., Kumar, Y., Zuhair, M., 2023. Prediction of significant wave height using machine learning and its application to extreme wave analysis. J. Earth Syst. Sci. 132 (2), 51.

Akhlaghi, Y.G., Aslansefat, K., Zhao, X., Sadati, S., Badiei, A., Xiao, X., Shittu, S., Fan, Y., Ma, X., 2021. Hourly performance forecast of a dew point cooler using explainable Artificial Intelligence and evolutionary optimisations by 2050. Appl. Energy 281, 116062.

Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115.

Awad, M., Khanna, R., 2015. Support vector regression. In: Efficient Learning Machines. Springer, pp. 67–80.

Barandela, R., Sánchez, J.S., Garcıa, V., Rangel, E., 2003. Strategies for learning in class imbalance problems. Pattern Recognit. 36 (3), 849–851.

Batista, G.E., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD Explor. Newsl. 6 (1), 20–29.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Cagman, N., Enginoglu, S., Citak, F., 2011. Fuzzy soft set theory and its applications. Iran. J. Fuzzy Syst. 8 (3), 137–147.

Chen, J., Zeng, G.-Q., Zhou, W., Du, W., Lu, K.-D., 2018. Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization. Energy Convers. Manage. 165, 681–695.

Chung, J., Gulcehre, C., Cho, K., Bengio, Y., 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Demetriou, D., Michailides, C., Papanastasiou, G., Onoufriou, T., 2021. Coastal zone significant wave height prediction by supervised machine learning classification algorithms. Ocean Eng. 221, 108592.

Ding, J., Deng, F., Liu, Q., Wang, J., 2023. Regional forecasting of significant wave height and mean wave period using EOF-EEMD-SCINet hybrid model. Appl. Ocean Res. 136, 103582.

Dixit, P., Londhe, S., 2016. Prediction of extreme wave heights using neuro wavelet technique. Appl. Ocean Res. 58, 241–252.

Draper, N.R., Smith, H., 1998. Applied regression analysis, vol. 326. John Wiley & Sons.

Durán-Rosal, A., Fernández, J., Gutiérrez, P., Hervás-Martínez, C., 2017. Detection and prediction of segments containing extreme significant wave heights. Ocean Eng. 142, 268–279.

Dysthe, K., Krogstad, H.E., Müller, P., 2008. Oceanic rogue waves. Annu. Rev. Fluid Mech. 40, 287–310.

Falcao, A.F.d.O., 2010. Wave energy utilization: A review of the technologies. Renew. Sustain. Energy Rev. 14 (3), 899–918.

Farahbod, S., Niknam, T., Mohammadi, M., Aghaei, J., Shojaeiyan, S., 2022. Probabilistic and deterministic wind speed prediction: Ensemble statistical deep regression network. IEEE Access.

Feng, Z., Hu, P., Li, S., Mo, D., 2022. Prediction of significant wave height in offshore china based on the machine learning method. J. Mar. Sci. Eng. 10 (6), 836.

Gao, S., Wang, Y., 2022. Explainable deep learning powered building risk assessment model for proactive hurricane response. Risk Anal..

Gao, Y., Wang, J., Zhang, X., Li, R., 2022. Ensemble wind speed prediction system based on envelope decomposition method and fuzzy inference evaluation of predictability. Appl. Soft Comput. 109010.

Gardner, M.W., Dorling, S.R., 1998. Artificial neural networks (the multilayer perceptron)-A review of applications in the atmospheric sciences. Atmos. Environ. 32, 2627–2636.

Gómez-Orellana, A.M., Fernández, J.C., Dorado-Moreno, M., Gutiérrez, P.A., Hervás-Martínez, C., 2021. Building suitable datasets for soft computing and machine learning techniques from meteorological data integration: A case study for predicting significant wave height and energy flux. Energies 14 (2), 468.

Gómez-Orellana, A., Guijo-Rubio, D., Gutiérrez, P., Hervás-Martínez, C., 2022. Simultaneous short-term significant wave height and energy flux prediction using zonal multi-task evolutionary artificial neural networks. Renew. Energy 184, 975–989.

Gómez-Orellana, A.M., Guijo-Rubio, D., Gutiérrez, P.A., Hervás-Martínez, C., Vargas, V.M., 2024. ORFEO: Ordinal classifier and regressor fusion for estimating an ordinal categorical target. Eng. Appl. Artif. Intell. 133, 108462.

Gómez-Orellana, A.M., Guijo-Rubio, D., Pérez-Aracil, J., Gutiérrez, P.A., Salcedo-Sanz, S., Hervás-Martínez, C., 2023. One month in advance prediction of air temperature from reanalysis data with eXplainable Artificial Intelligence techniques. Atmos. Res. 284, 106608.

Guijo-Rubio, D., Durán-Rosal, A.M., Gómez-Orellana, A.M., Fernández, J.C., 2023. An Evolutionary Artificial Neural Network approach for spatio-temporal wave height time series reconstruction. Appl. Soft Comput. 146, 110647.

Guijo-Rubio, D., Gómez-Orellana, A.M., Gutiérrez, P.A., Hervás-Martínez, C., 2020. Short-and long-term energy flux prediction using multi-task evolutionary artificial neural networks. Ocean Eng. 216, 108089.

Güner, H.A.A., Yüksel, Y., Çevik, E.Ö., 2013. Estimation of wave parameters based on nearshore wind–wave correlations. Ocean Eng. 63, 52–62.

Hansom, J.D., Switzer, A.D., Pile, J., 2015. Extreme waves: Causes, characteristics, and impact on coastal environments and society. In: Coastal and Marine Hazards, Risks, and Disasters. Elsevier, pp. 307–334.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.

Huang, G.B., Zhu, Q.Y., Siew, C.K., 2006. Extreme learning machine: Theory and applications. Neurocomputing 70, 489–501.

Hüsken, M., Stagge, P., 2003. Recurrent neural networks for time series classification. Neurocomputing 50, 223–235.

Ibarra-Berastegi, G., Saénz, J., Esnaola, G., Ezcurra, A., Ulazia, A., 2015. Short-term forecasting of the wave energy flux: Analogues, random forests, and physics-based models. Ocean Eng. 104, 530–539.

Ilic, I., Görgülü, B., Cevik, M., Baydoğan, M.G., 2021. Explainable boosted linear regression for time series forecasting. Pattern Recognit. 120, 108144.

Iong, D., Chen, Y., Toth, G., Zou, S., Pulkkinen, T., Ren, J., Camporeale, E., Gombosi, T., 2022. New findings from explainable SYM-H forecasting using gradient boosting machines. Space Weather 20 (8), e2021SW002928.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.-A., Petitjean, F., 2020. Inceptiontime: Finding alexnet for time series classification. Data Min. Knowl. Discov. 34 (6), 1936–1962.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K.C., Ropelewski, C., Wang, J., Jenne, R., Joseph, D., 1996. The NCEP/NCAR 40-year reanalysis project. Bull. Am. Meteorol. Soc. 77 (3), 437–471.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.

Kistler, R., Collins, W., Saha, S., White, G., Woollen, J., Kalnay, E., Chelliah, M., Ebisuzaki, W., Kanamitsu, M., Kousky, V., van den Dool, H., Jenne, R., Fiorino, M., 2001. The NCEP–NCAR 50–year reanalysis: Monthly means CD–ROM and documentation. Bull. Am. Meteorol. Soc. 82 (2), 247–267.

Kumar, N.K., Savitha, R., Al Mamun, A., 2018. Ocean wave height prediction using ensemble of extreme learning machine. Neurocomputing 277, 12–20.

Liu, W., Wang, C., Li, Y., Liu, Y., Huang, K., 2021. Ensemble forecasting for product futures prices using variational mode decomposition and artificial neural networks. Chaos Solitons Fractals 146, 110822.

Loh, W.-Y., 2011. Classification and regression trees. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 1 (1), 14–23.

Minuzzi, F.C., Farina, L., 2024. Artificial neural networks ensemble methodology to predict significant wave height. Ocean Eng. 300, 117479.

Molodtsov, D., 1999. Soft set theory—first results. Comput. Math. Appl. 37 (4–5), 19–31.

Mudronja, L., Matić, P., Katalinić, M., 2017. Data-based modelling of significant wave height in the Adriatic sea. Trans. Marit. Sci. 6 (01), 5–13.

National Data Buoy Center, 2023. National oceanic and atmospheric administration of the USA (NOAA). http://www.ndbc.noaa.gov/. (Accessed 13 December 2023).

Peláez-Rodríguez, C., Pérez-Aracil, J., Fister, D., Prieto-Godino, L., Deo, R., Salcedo-Sanz, S., 2022. A hierarchical classification/regression algorithm for improving extreme wind speed events prediction. Renew. Energy 201, 157–178.

Peláez-Rodríguez, C., Pérez-Aracil, J., Marina, C., Prieto-Godino, L., Casanova-Mateo, C., Gutiérrez, P., Salcedo-Sanz, S., 2024. A general explicable forecasting framework for weather events based on ordinal classification and inductive rules combined with fuzzy logic. Knowl.-Based Syst. 111556.

Peláez-Rodríguez, C., Pérez-Aracil, J., Prieto-Godino, L., Ghimire, S., Deo, R., Salcedo-Sanz, S., 2023. A fuzzy-based cascade ensemble model for improving extreme wind speeds prediction. J. Wind Eng. Ind. Aerodyn. 240, 105507.

Pérez-Aracil, J., Camacho-Gómez, C., Lorente-Ramos, E., Marina, C.M., Cornejo-Bueno, L.M., Salcedo-Sanz, S., 2023. New probabilistic, dynamic multi-method ensembles for optimization based on the CRO-SL. Mathematics 11 (7), 1666.

Pérez-Aracil, J., Camacho-Gómez, C., Lorente-Ramos, E., Marina, C.M., Salcedo-Sanz, S., 2022. New probabilistic-dynamic multi-method ensembles for optimization based on the CRO-SL. arXiv preprint arXiv:2212.00742.

Petrov, V., Soares, C.G., Gotovac, H., 2013. Prediction of extreme significant wave heights using maximum entropy. Coast. Eng. 74, 1–10.

Prado, F., Minutolo, M.C., Kristjanpoller, W., 2020. Forecasting based on an ensemble autoregressive moving average-adaptive neuro-fuzzy inference system–neural network-genetic algorithm framework. Energy 197, 117159.

Prechelt, L., et al., 1994. Proben1: A Set of Neural Network Benchmark Problems and Benchmarking Rules. Technical Report, Citeseer, Fakultät für Informatik, Universität Karlsruhe.

Ren, Y., Zhang, L., Suganthan, P.N., 2016. Ensemble classification and regression-recent developments, applications and future directions. IEEE Comput. Intell. Mag. 11 (1), 41–53.

Rueda, A., Camus, P., Méndez, F.J., Tomás, A., Luceño, A., 2016. An extreme value model for maximum wave heights based on weather types. J. Geophys. Res.: Oceans 121 (2), 1262–1273.

Saeed Far, S., Abd. Wahab, A.K., 2016. Evaluation of peaks-over-threshold method. Ocean Sci. Discuss. 2016, 1–25.

Salcedo-Sanz, S., Camacho-Gómez, C., Molina, D., Herrera, F., 2016. A coral reefs optimization algorithm with substrate layers and local search for large scale global optimization. In: 2016 IEEE Congress on Evolutionary Computation. CEC, IEEE, pp. 3574–3581.

Salcedo-Sanz, S., Del Ser, J., Landa-Torres, I., Gil-López, S., Portilla-Figueras, J., 2014. The coral reefs optimization algorithm: a novel metaheuristic for efficiently solving optimization problems. Sci. World J. 2014.

Samayam, S., Laface, V., Annamalaisamy, S.S., Arena, F., Vallam, S., Gavrilovich, P.V., 2017. Assessment of reliability of extreme wave height prediction models. Nat. Hazards Earth Syst. Sci. 17 (3), 409–421.

Shamshirband, S., Mosavi, A., Rabczuk, T., Nabipour, N., Chau, K.-w., 2020. Prediction of significant wave height; comparison between nested grid numerical model, and machine learning models of artificial neural networks, extreme learning and support vector machines. Eng. Appl. Comput. Fluid Mech. 14 (1), 805–817.

Shi, Q., Katuwal, R., Suganthan, P.N., Tanveer, M., 2021. Random vector functional link neural network based ensemble deep learning. Pattern Recognit. 117, 107978.
Sideratos, G., Ikonomopoulos, A., Hatziargyriou, N.D., 2020. A novel fuzzy-based ensemble model for load forecasting using hybrid deep neural networks. Electr. Power Syst. Res. 178, 106025.
Sushanth, K., Mishra, A., Mukhopadhyay, P., Singh, R., 2023. Real-time streamflow forecasting in a reservoir-regulated river basin using explainable machine learning and conceptual reservoir module. Sci. Total Environ. 861, 160680.
Ti, Z., Li, Y., Qin, S., 2020. Numerical approach of interaction between wave and flexible bridge pier with arbitrary cross section based on boundary element method. J. Bridge Eng. 25 (11), 04020095.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.
Toms, B.A., Barnes, E.A., Ebert-Uphoff, I., 2020. Physically interpretable neural networks for the geosciences: Applications to earth system variability. J. Adv. Modelling Earth Syst. 12 (9), e2019MS002002.
Torgo, L., Ribeiro, R.P., Pfahringer, B., Branco, P., 2013. Smote for regression. In: Portuguese Conference on Artificial Intelligence. Springer, pp. 378–389.
Viselli, A.M., Forristall, G.Z., Pearce, B.R., Dagher, H.J., 2015. Estimation of extreme wave and wind design parameters for offshore wind turbines in the Gulf of Maine using a POT method. Ocean Eng. 104, 649–658.
Wang, H., Hu, D., 2005. Comparison of SVM and LS-SVM for regression. In: 2005 International Conference on Neural Networks and Brain, Vol. 1. IEEE, pp. 279–283.
Wang, Z., Yan, W., Oates, T., 2017. Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International Joint Conference on Neural Networks. IJCNN, IEEE, pp. 1578–1585.
Wu, G., Mallipeddi, R., Suganthan, P.N., 2019. Ensemble strategies for population-based optimization algorithms–A survey. Swarm Evol. Comput. 44, 695–711.
Zhao, J., Mao, X., Chen, L., 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomed. Signal Process. Control 47, 312–323.
Zilong, T., Wei, D.X., 2022. Layout optimization of offshore wind farm considering spatially inhomogeneous wave loads. Appl. Energy 306, 117947.
Zilong, T., Yubing, S., Xiaowei, D., 2022. Spatial-temporal wave height forecast using deep learning and public reanalysis dataset. Appl. Energy 326, 120027.