Contents lists available at ScienceDirect

Applied Soft Computing

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

An Evolutionary Artificial Neural Network approach for spatio-temporal wave height time series reconstruction

David Guijo-Rubio [a,b,*,1], Antonio M. Durán-Rosal [c,1], Antonio M. Gómez-Orellana [a], Juan C. Fernández [a]

[a] Department of Computer Science and Numerical Analysis, University of Córdoba, Córdoba, Spain
[b] School of Computing Sciences, University of East Anglia, Norwich, United Kingdom
[c] Department of Quantitative Methods, Universidad Loyola Andalucía, c/ Escritor Aguayo, 4, Córdoba, Spain

ARTICLE INFO

Article history:
Received 19 September 2022
Received in revised form 6 June 2023
Accepted 10 July 2023
Available online 20 July 2023

Keywords:
Time series reconstruction
Marine engineering
Wave height reconstruction
Evolutionary Artificial Neural Networks
Ocean buoys

ABSTRACT

This paper proposes a novel methodology for recovering missing time series data, a crucial task for subsequent Machine Learning (ML) analyses. The methodology is specifically applied to Significant Wave Height (SWH) time series in the field of marine engineering. The proposed approach involves two phases. Firstly, the SWH time series for each buoy is independently reconstructed using three transfer function models: regression-based, correlation-based, and distance-based. The distance-based transfer function exhibits the best overall performance. Secondly, Evolutionary Artificial Neural Networks (EANNs) are utilised for the final recovery of each time series, using as inputs highly correlated buoys that have been intermediately recovered. The EANNs are evolved considering two metrics, the novel squared error relevance area, which balances the importance of extreme and around-mean values, and the well-known mean squared error. The study considers SWH time series data from 15 buoys in two coastal zones in the United States. The results demonstrate that the distance-based transfer function is generally the best transfer function, and that EANNs outperform a range of state-of-the-art ML techniques in 12 out of the 15 buoys, with a number of connections comparable to linear models. Furthermore, the proposed methodology outperforms the two most popular approaches for time series reconstruction, BRITS and SAITS, for all buoys except one. Therefore, the proposed methodology provides a promising approach, which may be applied to time series from other fields, such as wind or solar energy farms in the field of green energy.

1. Introduction

Significant Wave Height (SWH) has attracted a considerable worldwide interest because of its key role in marine engineering and resource development [1], shipping and maritime transport [2], fishery and aquaculture [3], and wave energy prediction [4], among others.

In this regard, oceanographic buoys, maintained by agencies of different countries and deployed around oceans and seas, are used to measure SWH, as well as other wave-related parameters. This is the case of the National Data Buoy Center (NDBC) and the National Oceanic and Atmospheric Administration (NOAA) [5]. They use hydrographic stations and ocean buoys equipped with special sensors to collect environmental data from coastal regions of the United States of America (USA). Although there are buoys located close to the shore, the vast majority are deployed in remote or inaccessible locations for long time periods [6]. Besides, extreme weather conditions such as cold, storms or cyclones [7], as well as unexpected events such as accidents, sensor failures or technical maintenance [8], cause buoys to be inoperative during different periods of time. This implies the existence of two sorts of missing values in the SWH time series [9]: (1) extended periods of time without collecting data, and (2) intermittent missing values. Although they can be interpreted by some methods, such as clustering [10], it is advisable to recover them so that methods not capable of dealing with missing values can be applied.

As in other areas, for example, in climatology with the amount of rainfall [11] or the temperature [12,13], the reconstruction of SWH missing values, in this case, is considered an essential pre-processing phase, given that SWH analysis and prediction tasks have a high impact on human and economic activities. For example, maritime transport management [14], the influence of oceanographic parameters on fishing [15], or the design, planning

[*] Corresponding author at: Department of Computer Science and Numerical Analysis, University of Córdoba, Córdoba, Spain.
E-mail addresses: dguijo@uco.es (D. Guijo-Rubio), amduran@uloyola.es (A.M. Durán-Rosal), am.gomez@uco.es (A.M. Gómez-Orellana), jfcaballero@uco.es (J.C. Fernández).
[1] D. Guijo-Rubio and A.M. Durán-Rosal contributed equally to this work.

https://doi.org/10.1016/j.asoc.2023.110647

and maintenance of offshore structures [16]. Moreover, SWH is very relevant in some climatic processes [17] and in forecasting oceanic cyclones, earthquakes and tsunamis [18]. However, nowadays, the energy management being harnessed is one of the most paramount concerns of predicting the SWH, and consequently the prediction of wave features. SWH has become one of the critical factors in wave energy determination as presented in a large number of works and research in recent years [19–23]. Related to this, the exploration of ocean energy resources has demonstrated a clear potential for sustainable growth, becoming a major issue for renewable energies [24]. In fact, there is an enormous need to develop new methodologies capable of efficiently exploiting the natural resources available in our environment. In this sense, one of the most promising clean energy sources in recent years is marine energy [25]. Consequently, the reconstruction of missing SWH time series has emerged as a key topic in the ocean research field [26].

Despite extensive research, SWH time series reconstruction remains a challenging problem and a growing area of research due to its crucial role in marine engineering and renewable energy estimation. In this sense, efforts are addressed to explore new approaches aimed at overcoming the flaws of the state-of-the-art methods.

Therefore, the purpose of this work is to propose a novel methodology for recovering missing values in SWH time series collected by ocean buoys located at different geographical locations. Specifically, the methodology is divided into two phases: (1) the missing values are intermediately recovered by applying different Transfer Function (TF) models and considering the available data from neighbouring buoys, and (2) the intermediately recovered data is used by Evolutionary Artificial Neural Networks (EANNs) to perform the final reconstruction of the SWH time series. One of the main innovations of this novel spatio-temporal approach is that it benefits from retrieving information not only temporally, but also geographically, as it takes into account the proximity to other buoys. The main contributions of this paper are:

- TF models perform an intermediate recovery (first phase) which is subsequently used as input for EANNs to perform the final reconstruction (second phase).
- This paper proposes a new distance-based TF model that improves the intermediate recovery for several of the buoys considered in this study.
- One of the main drawbacks of most state-of-the-art methods is that they attempt to minimise the average error made in estimating missing SWH values (i.e. the Mean Squared Error (MSE)). This results in a much larger error in the highest waves, also known as extreme waves. In this way, this paper proposes the use of a recently developed metric that pays attention to the extreme values without leaving aside the around-mean values. This metric is known as Squared Error-Relevance Area (SERA) [27], and aims to balance the attention paid to both extreme and around-mean values, regardless of the quantity of one type or another. Hence, for the extreme values, models optimised by SERA achieve more accurate estimations than models optimised by MSE.
- In the second phase, the performance of three EANN architectures is analysed regarding the type of basis function for neurons in the hidden layer: sigmoidal unit, product unit, as well as the combination of both types (hybridisation) with the idea of taking advantage of both to improve the final reconstruction of the SWH time series.
- Given the stochastic nature of the proposed EANN technique, for each SWH time series reconstruction, 30 runs are performed to analyse the average results obtained, rather than analysing only the best one as is done in other works in the literature.
- The feasibility and robustness of the proposed methodology are assessed considering two coastal zones of the USA, namely the Northeast Coast and the Gulf of Alaska, with a total of 15 buoys.
- The results achieved by the EANN technique are compared against other 7 ML methods that have been specifically adapted for the problem tackled. Also, a second comparison is carried out against the 2 state-of-the-art approaches in the time series imputation field.
- The proposed methodology has been designed in such a way that it can be applied to related problems, as is the case of wind and solar energy farms in the green energies field, or to fog and rainfall detection systems at airports in the atmospheric area, among others.

The remainder of this paper is organised as follows: Section 2 presents related works in the literature. Section 3 details the SWH time series employed. Section 4 explains the two-phase methodology proposed for the recovery of missing SWH time series values. Section 5 describes the experimental design, shows the results obtained and provides a detailed discussion comparing the methodology with other state-of-the-art algorithms. Finally, Section 6 concludes the paper and includes future work.

2. Related works

In recent decades, many approaches have been published in the literature aiming to recover missing SWH time series data. On the one hand, traditional and statistically-based methods were often used for this purpose. For example, traditional proposals using random sampling [28] or Monte Carlo methods [29] were initially presented to recover SWH time series. Regarding statistical techniques, in [30], the authors proposed a method based on the AutoRegressive Moving Average (ARMA) model with a prior transformation of the input data to reconstruct a SWH time series collected at the Portuguese coast. Similarly, a methodology using an ARMA model based on non-stationary modelling of long-term SWH time series was presented in [31], in which missing values were recovered at the level of the uncorrelated residuals.

On the other hand, there has been a growing trend in the use of Machine Learning (ML) approaches for the reconstruction of SWH time series. There is a myriad of ML techniques capable of achieving excellent results, such as Gaussian Process Regression (GPR) [32], Support Vector Machine (SVM) [33] or Extreme Learning Machine (ELM) [34], among others. However, Artificial Neural Network (ANN) models [35] as for being the most widely used technique to address SWH missing values recovery. In [36], an ANN model was proposed for the spatio-temporal analysis of SWH time series collected by a network of buoys, aiming to determine the best way to recover the gaps in SWH time series. For this, the performance of ANNs was compared with observed data collected by stations located at the Ionian and Adriatic Seas, showing the reliability of this model. Silva-Ramírez et al. in [37], proposed two imputation approaches: a single imputation technique based on a Multilayer Perceptron (MLP) model trained by means of different learning rules and a multiple imputation technique based on combining MLP with k-Nearest Neighbours (kNN). In [38], the authors demonstrated that the proposed Elman-type recurrent ANN models, trained with both steepest descent with impulse and conjugate gradients methods, outperformed the MLP model. In [39], the feasibility of three different ANN architectures for wave data supplementation was assessed using measurements collected near the Tasmanian coast. Moreover, in recent years, the emergence of SWH reconstruction and prediction works using deep ANNs has exponentially increased [40]. In this regard, the use of recurrent ANNs and their more widespread type, Long-Short Term Memory (LSTM) [41], has been satisfactorily applied in this field: in [42], the authors presented a framework using LSTM models for the reconstruction of coastal sea levels time series in Korea, in [43], bathymetric data was used to improve the performance of LSTM models in the West Coast of the USA, and Pirhooshyaran and Snyder [44], also through LSTM, demonstrated that for a large number of target features, deeper structures are able to improve the performance of other ML techniques.

Therefore, ANNs have proven to be effective and accurate when tackling the problem addressed in this work. However, they show some constraints, such as the determination of the most suitable architecture (number of neurons in hidden layer and connections) for the problem being solved or local minima issues during the training phase. In this sense, Evolutionary Computation (EC) is an excellent technique to optimise the architecture of ANNs and enhance their performance. EC comprises a set of optimisation algorithms which apply biologically inspired mechanisms, including reproduction, mutation, natural selection and survival of the fittest, among others [45]. Some of these algorithms are Genetic Algorithm (GA), Particle Swarm Optimisation (PSO), Coral Reefs Optimisation (CRO) and Evolutionary Algorithm (EA), to name a few [46].

For instance, in [47], the authors proposed a hybrid ELM combined with a GA for tuning the parameters to solve the reconstruction of SWH time series at the Caribbean Sea and West Atlantic. A similar hybrid ELM approach, also trained by a GA, was analysed to find the best subset of features [24], such methodology was proposed to estimate the SWH and energy flux at the Western Coast of the USA. In [48], PSO was employed to train ANNs for SWH time series estimation at New Mangalore Port in India. In [49], a standard Back-Propagation (BP) ANN was improved using a Mind EA to predict ocean wave heights at the Bohai Sea and the Yellow Sea, showing that the novel approach was more suitable when compared to its BP and GA version. A methodology combining ANNs optimised by means of an EA with linear models was presented in [26] for recovering missing values in SWH time series. Recently, in [50], the authors developed a novel approach to simultaneously predict short-term SWH and energy flux at the South West Coast of the USA and at the Gulf of Alaska, using Multi-Task ANNs trained by an EA. Hence, the combination of ANNs and EC has demonstrated its great relevance and high performance in tackling SWH reconstruction and related problems in this field of application, which has been gaining momentum in recent years.

3. Data description

The data used in this study has been obtained from the NDBC [51]. Specifically, NDBC records meteorological and oceanographic measurements regarding the marine environment using buoys deployed along coastal regions of the USA. One of the measurements collected by the sensors installed in the buoys is the Significant Wave Height (SWH), which is the object under study in this work. In addition to the SWH, the sensors also record other environmental observations, such as wind speed and air temperature.

This work considers the SWH time series of 15 buoys located at two relevant coastal zones of the USA. The following Sections describe this data.

3.1. Significant wave height

As aforementioned, when sensors are inoperative, none of these measurements can be collected. Even though the reconstruction of missing values in any of these measurements is of great interest, the reason behind choosing the SWH is its enormous impact on highly relevant areas, such as marine engineering or renewable energy production. At this point, it is important to note how SWH is defined. On the one hand, if SWH is defined in the temporal domain, $H_{1/3}$ is noted and defined as the mean height of the highest third of the wave heights, measured from the time series of the free surface by the upward or downward crossover. On the other hand, if the definition is in the frequency domain, H_{m0} is noted and defined from the frequency spectrum. However, in deep water, both measurements are less than 5% different, and they are usually confused with the generic term H_s. For this reason, although definitions of wave height are formally expressed, it is recommended to use the generic term H_s or simply SWH, which is defined as the mean in metres between the trough and the crest of the highest third of all wave heights during a 20-minute sampling period [52]. Specifically, this sampling period is the one used by the processor on board of the buoys to obtain the SWH time series considered in this study [53].

3.2. Zones subject to research

Two different coastal zones of the USA have been selected for this work: the Gulf of Alaska and the Northeast Coast, where 6 and 9 buoys have been considered, respectively. Fig. 1 shows the location of each of these 15 buoys: the Figure in the middle shows the geographical location of both coastal zones in the USA, whereas the upper and the lower Figures show the specific location of the buoys at the Gulf of Alaska and at the Northeast Coast, respectively.

These two zones of the USA have been considered given the entirely different environmental characteristics they present. As can be seen in Fig. 1, the Gulf of Alaska is located at the West Coast of the USA, whereas the Northeast Coast is located, as its name suggests, at the East Coast. Moreover, the Gulf of Alaska is closer to the North Pole, whereas the Northeast Coast is closer to the Equator. Table 1 presents the geographical location and water depth of each buoy. From this Table 1, it is interesting to remark the differences in water depth between both zones: the water depth in the Northeast Coast is, in general, very low, the deepest being at 185 m (buoy 44027), in comparison with the Gulf of Alaska, where the minimum water depth is 192 m (buoy 46076) and for the half of the buoys is higher than 3500 m.

3.3. SWH time series

For each buoy, its corresponding SWH time series has been recorded 4 times daily (i.e. 6 h resolution) from year 2013 to 2018, resulting in a total of 8764 values per time series. Table 2 summarises the number of training, testing and missing values for each SWH time series. These values have been obtained separately for each buoy in the following way: first of all, the missing values have been identified, then, the remaining ones (i.e. the non-missing values) have been randomly divided into the training and testing sets (80% and 20% of the non-missing values, respectively) with the constraint that the 20% of the testing values must be consecutively selected from the time series, that is, not having any gap. Thus, the first sequence having a 20% of the non-missing values consecutively defines the testing set. Because of the randomness and the availability of a 20% consecutive non-missing values, the time instants for training and testing may vary from one buoy to the others.

To better understand the procedure described and the different sets, Fig. 2 shows two SWH time series corresponding to the 46061 and the 44005 buoys of the Gulf of Alaska and the Northeast Coast, respectively. Values in green colour belong to the training set, whereas those coloured in red belong to the testing

Table 1
Geographical location and water depth of the buoys.

	Buoy	Geographical location (latitude N, longitude W)	Water depth (m)
Gulf of Alaska	46001	(56.232N, 147.949W)	4054
	46061	(60.238N, 146.833W)	222
	46076	(59.471N, 148.009W)	192
	46078	(55.556N, 152.582W)	5380
	46082	(59.681N, 143.372W)	300
	46085	(55.883N, 142.882W)	3721
	Buoy	Geographical location (latitude N, longitude W)	Water depth (m)
Northeast Coast	44005	(43.201N, 69.127W)	177
	44008	(40.498N, 69.251W)	69
	44011	(41.093N, 66.562W)	91
	44013	(42.346N, 70.651W)	65
	44020	(41.493N, 70.279W)	14
	44025	(40.251N, 73.164W)	36
	44027	(44.283N, 67.300W)	185
	44065	(40.369N, 73.703W)	25
	44066	(39.618N, 72.644W)	78

Table 2
Description of the SWH time series of each buoy.

	Buoy	# training values	# testing values	# missing values
Gulf of Alaska	46001	8764 (100.000%)	–	–
	46061	5374 (61.319%)	1343 (15.324%)	2047 (23.357%)
	46076	6086 (69.443%)	1522 (17.366%)	1156 (13.190%)
	46078	5138 (58.626%)	1284 (14.651%)	2342 (26.723%)
	46082	4258 (48.585%)	1065 (12.152%)	3441 (39.263%)
	46085	5148 (58.740%)	1287 (14.685%)	2329 (26.575%)
	Buoy	# training values	# testing values	# missing values
Northeast Coast	44005	4429 (50.536%)	1107 (12.631%)	3228 (36.832%)
	44008	3850 (43.930%)	962 (10.977%)	3952 (45.094%)
	44011	3328 (37.974%)	832 (9.493%)	4604 (52.533%)
	44013	8764 (100.000%)	–	–
	44020	6862 (78.298%)	1716 (19.580%)	186 (2.122%)
	44025	5827 (66.488%)	1457 (16.625%)	1480 (16.887%)
	44027	5940 (67.777%)	1485 (16.944%)	1339 (15.278%)
	44065	5798 (66.157%)	1450 (16.545%)	1516 (17.298%)
	44066	5755 (65.666%)	1439 (16.419%)	1570 (17.914%)

set. Moreover, gaps indicate the presence of missing values as intuitive. As aforementioned, missing values can be found in two different ways: (1) intermittent periods of time, and (2) long consecutive ones. An example of the first type is shown in the buoy 46061, wherein the time period from Jan-2013 to Apr-2014 of the SWH time series, a high amount of the missing values is intermingled with data. On the other hand, the second behaviour is shown for the buoy 44005, where the whole time period from Jan-2013 to Mar-2014 of the SWH time series is missing. It can also be observed, as mentioned above, that the testing set (coloured in red) vary from one buoy to the others, and they do not contain gaps.

As can be seen in Table 2, the percentage of missing values varies from 2.122% (buoy 44020) to 52.533% (buoy 44011), meaning that recovering the first buoy should be easier than recovering the second one. In this regard, it is important to specify that the concept of complexity is not only associated with the amount of missing data existing in the time series but also with its dynamics. Besides, it is worthy of mention that for the proposed regression-based TF, at least one buoy without missing values is required per zone since it is considered the starting point from which the missing values of neighbouring buoys are recovered. In this sense, buoys 46001 and 44013 belonging to the Gulf of Alaska and the Northeast Coast, respectively, are complete, as can be observed in Table 2.

Therefore, the proposed approach aims to recover both sorts of missing values in order to provide an accurate reconstruction of the missing data so that SWH time series can be used in subsequent tasks, such as prediction, classification or clustering, among others.

4. Methodology

The proposed methodology is divided into two phases. First, an intermediate reconstruction of the missing values in the SWH time series is performed by applying different Transfer Function (TF) models. These TFs consider available data from neighbouring buoys and select the one achieving the lowest error. For readability purposes, this first phase will be named as intermediate recovery. After that, in the second phase, known as final recovery, Evolutionary Artificial Neural Network (EANN) models are employed to perform the final and definitive reconstruction. For this, the rest of the intermediately recovered buoys are used as input. It is worth mentioning that the first phase allows the use of any ML technique, such as EANNs, which cannot be applied when missing data across time series do not coincide in time, i.e., the missing data is found at different time/points of the time series. Fig. 3 summarises the procedure described above. The following Sections detail both phases, including specific flowcharts per phase.

4.1. Phase 1: intermediate recovery

This first phase carries out an intermediate recovery of the SWH time series, which is performed using only data available

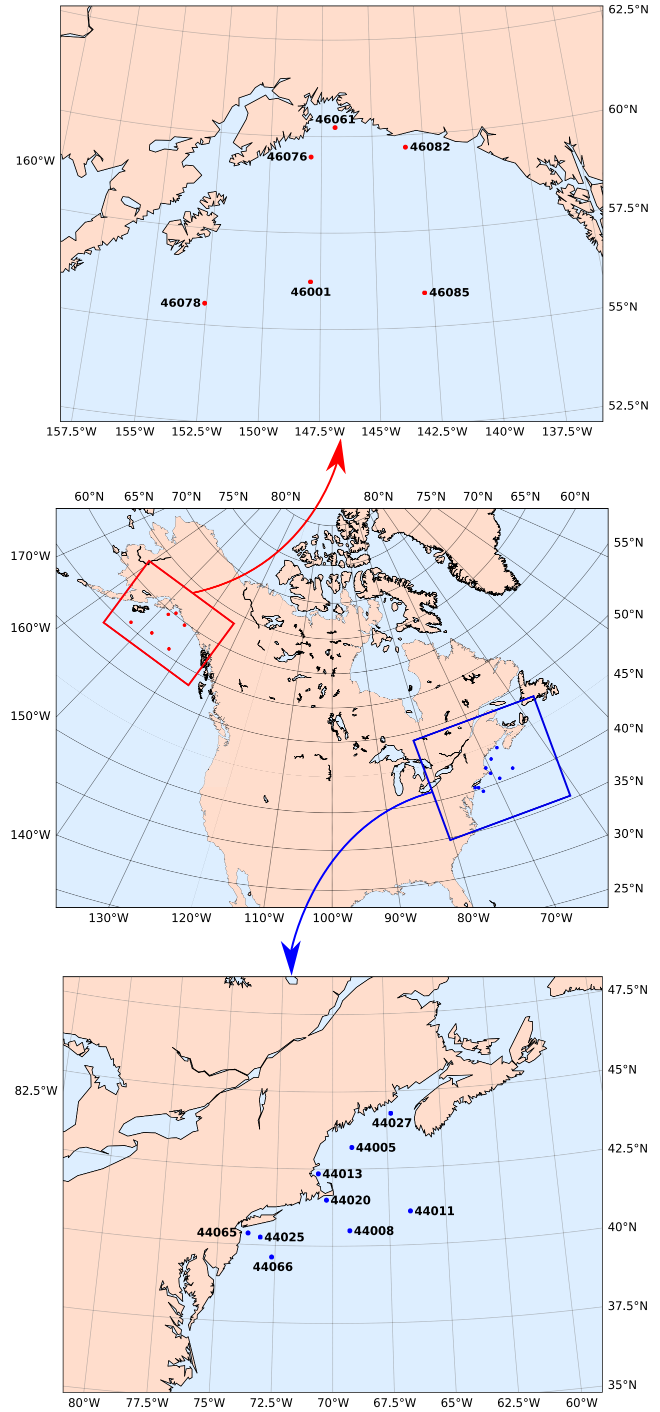

Fig. 1. Geographical location of the two zones under study (middle) and specific buoys location at the Gulf of Alaska (upper) and at the Northeast Coast (lower).

in the neighbouring buoys. For this, three completely independent TF models have been adapted and developed: regression-based, correlation-based and distance-based models. These TFs have the advantage that they do not require training, unlike ML techniques, which cannot be applied at this point, as the available information is not equivalent for all buoys. Nevertheless, TF models can be straightforwardly applied to the data available in the neighbouring buoys. It is important to mention that the regression-based TF is applied in a time series-wise manner, i.e. one regression model is built using two SWH time series. For this TF, at least one complete SWH time series has to be available. On the other hand, correlation and distance-based TFs are point-wisely applied. In other words, the information available is considered for each time instant. Hence, even though a complete SWH time series is not needed, it is required to have information from at least one buoy for each time instant. The three TF models, namely regression, correlation and distance-based TFs, are detailed in the following Sections. Fig. 4 graphically summarises the main steps of this first phase.

4.1.1. Regression-based model

This method assumes that a complete time series already exists among all those available in the buoy grid, i.e. the buoys belonging to the same coastal zone. In an iterative process, for each incomplete time series, its correlation with respect to those with available data is computed. It is worthy of mention that only buoys with available values at same time instants as the buoy being recovered are considered for computing the correlation, and that, at least, one correlation value is expected, since it is assumed that one buoy is complete. The time series with a correlation above a threshold α are the input variables of a regression model for the time series of the buoy to be recovered at that time. Thus, at each iteration, those SWH time series whose correlation with respect to the complete buoys is above such threshold are intermediately reconstructed. It is important to mention that when a time series is reconstructed, it is considered complete for the next iteration and can be used for recovering the time series of other buoys (if the correlation is above the α threshold). Furthermore, in each iteration, the α threshold is increased to compensate for the error incurred in the reconstruction process. The regression model is defined as follows:

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i X_{it}, \qquad t = 1, \ldots, N, \tag{1}$$

where Y_t is the value to be recovered at time instant t, β_i is the coefficient for the ith SWH time series considered as a predictor variable (X_i), k is the number of correlated buoys (those whose correlation is above the α threshold), and N is the length of the SWH time series. In this work, the least-squares method has been used to solve the regression problem so that:

$$\beta_i = \frac{-A_{1,i+1}}{A_{1,1}}, \qquad i = 1, \ldots, k, \tag{2}$$

where $A_{i,j}$ is the adjoint (i, j) of the variances and covariances matrix Σ. And therefore:

$$\beta_0 = \bar{Y} - \sum_{i=1}^{k} \beta_i \bar{X}_i, \tag{3}$$

being $\bar{Y}$ and $\bar{X}_i$, the average value of Y and X_i, respectively.

4.1.2. Correlation-based model

This method is based on recovering the missing values of the time series of each buoy by weighting the similarity with respect to the time series of the neighbouring buoys (remaining buoys of the same coastal zone). In this case, as the information is retrieved point-wisely, instead of needing to have a buoy with complete data, having information from at least one of the neighbouring buoys for the time stamp being reconstructed is sufficient. Thus, this aspect is easier to be fulfilled as having one buoy with complete data may be difficult to find in some areas. Specifically, missing values at time instant t to be recovered are weighted by a factor λ that measures the importance contributed by the same time instant t available at the neighbouring buoys. In this

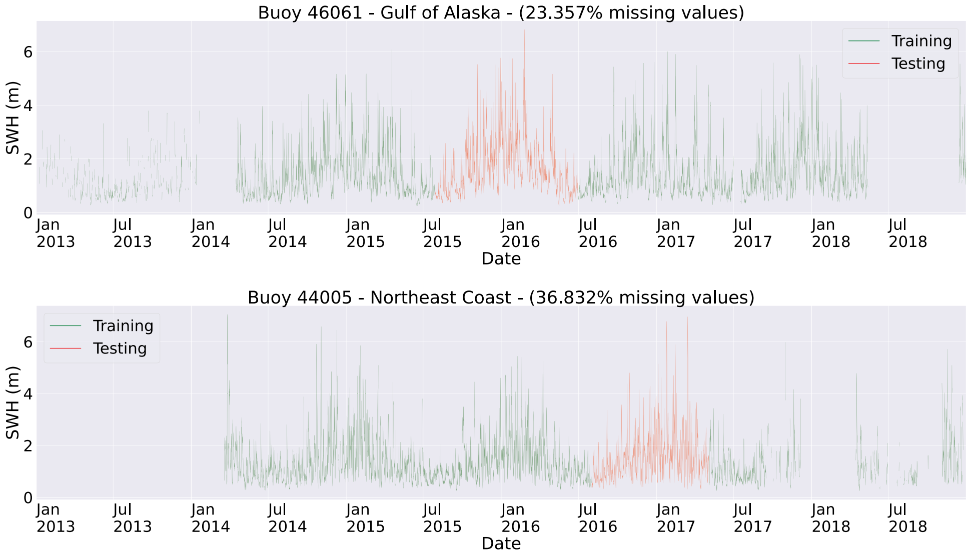

Fig. 2. SWH time series of the buoys 46061 and 44005 of the Gulf of Alaska and the Northeast Coast, respectively.

way, the time instant t of a SWH time series Y is reconstructed as follows:

$$Y_t = \sum_{i=1}^{k} \lambda_i \frac{\bar{Y} X_{it}}{\bar{X_i}}, \tag{4}$$

being k the number of SWH time series with data available at time instant t, $\bar{Y}$ and $\bar{X_i}$ are the mean values of Y and X_i, respectively, X_{it} is the value of X_i at time instant t, and λ_i is a multiplicative factor allowing a greater weight to buoys with higher correlation, and it is expressed as follows:

$$\lambda_i = \frac{\rho_{YX_i}}{\sum_{j=1}^{k} \rho_{YX_j}}, \qquad i = 1, \ldots, k, \tag{5}$$

being ρ_{YX_i} the Pearson correlation coefficient between Y and X_i. Hence, values of correlated buoys are paid more attention.

Even though correlation plays an essential role in both TFs, it is important to differentiate how it is employed in each one. The regression-based TF indirectly uses the correlation to decide which buoys are considered input for the regression model. On the other hand, the correlation-based TF directly weights the neighbouring buoys' values using the correlation between them. Moreover, note that this method is much faster than the previous one, as all missing values for a given buoy can be recovered simultaneously.

4.1.3. *Distance-based model*

Similar to the previous correlation-based TF, this model recovers each buoy time series by weighting the time series of the neighbouring buoys as defined in Eq. (4). However, instead of the correlation, the multiplicative factor λ is computed by considering the distance between buoys. Hence, the main goal of this technique is to measure the similarity according to the geographical position of the buoys, i.e. it is supposed that close buoys should have similar behaviour.

To calculate the distance between buoys, the *Haversine* equation (also known as the great circle distance) [54] has been considered. It is defined as follows:

$$\begin{aligned} d(p_0, p_1) = \arccos(&\sin(lat_0) \cdot \sin(lat_1) \\ &\cdot \cos(lon_0 - lon_1) + \cos(lat_0) \\ &\cdot \cos(lat_1)), \end{aligned} \tag{6}$$

where p_0 is the geographical location of the buoy being recovered, p_1 is the location of the neighbouring buoy with available data at the time instant being processed, and lat and lon are the latitude and longitude of the buoys, respectively.

This distance is inverted and normalised as follows, in such a way that the greater the distance, the smaller the weight that the neighbouring buoy should have:

$$\lambda_i = \frac{d(p_0, p_i)}{\sum_{j=1}^{k} d(p_0, p_j)}, \qquad i = 1, \ldots, k, \tag{7}$$

being λ_i the weight for the ith buoy with available data for recovering a given missing value, and k is the number of neighbouring buoys with available data. In this case, for the computation of the multiplicative factor λ, no data from the neighbouring time series has been considered but their geographical positions (latitude and longitude of the buoys).

At this point, it is important to analyse the relationship between correlation and distance. It is assumed that when one feature increases, the other decreases and vice-versa. However, this assumption is not always true. Highly correlated buoys are supposed to be geographically close to each other. Nevertheless, it may happen that buoys located at the coast but very far away, one from the other, could share some similarity in their dynamics as their geographical accidents may be similar. On the opposite, closer buoys are supposed to be highly correlated. But this assumption may not always be met either. There could be a geographical accident or a vast change in the orography of the ocean happening in the middle of two close buoys causing very different wave conditions in each of them, and, as a consequence, the data collected by each buoy is unrelated.

4.2. *Phase 2: final recovery*

In this second phase, the final and definitive reconstruction of each SWH time series is carried out. For this, the intermediate reconstructions of SWH time series of the remaining buoys in the zone are used as input for a model. At this point, it is possible to apply complex ML techniques. These techniques could not be previously applied as they require a training step, which, in turn, could not be accomplished given that for the training time instants of a given SWH time series, missing values could be found in other buoys. However, once the intermediate reconstruction is

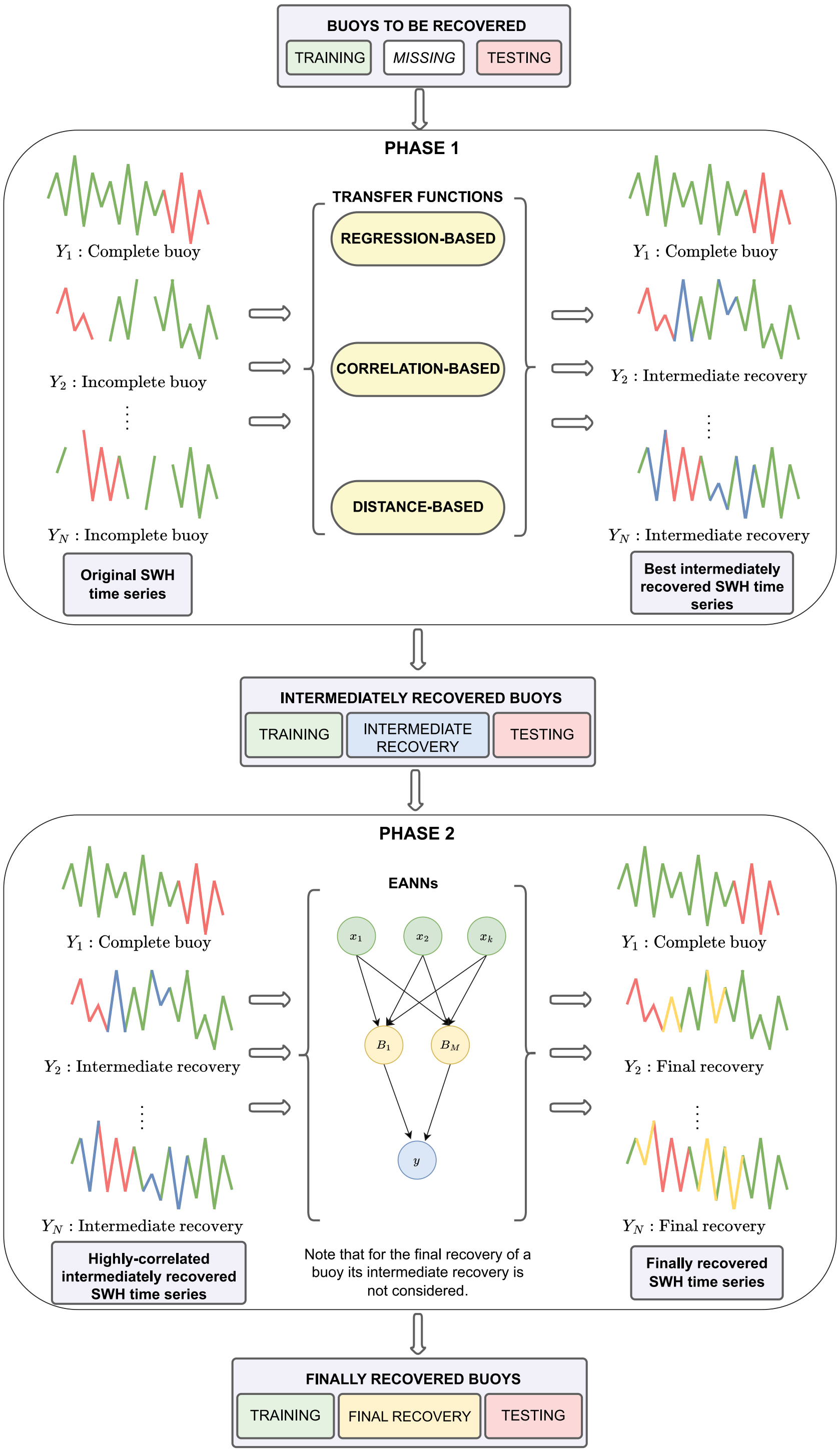

Fig. 3. Simplified flowchart of the complete methodology, where the intermediate recovery is performed in the first phase, using Transfer Functions, and the final recovery is carried out using Evolutionary Artificial Neural Network (EANN) models in the second phase.

performed, there are no missing values in the input SWH time series, hence, the training stage of these ML techniques can be carried out.

Furthermore, from all the buoys of a zone, only those highly-correlated SWH time series are used as inputs, i.e. those SWH time series with a correlation above the threshold α. Note that this threshold is the initial one and is not recomputed to avoid introducing bias from the intermediate recovery. In addition, as it was aforementioned, the testing set of each buoy remains invariant, and it may be used in the training stage of each of the remaining buoys, but it is not used in the buoy being recovered. Moreover, it is important to mention that its intermediate recovery is not considered as input for the final reconstruction of the SWH time series for a specific buoy. To proceed with, EANNs are proposed for this second phase to obtain more accurate estimations as they have been previously applied, achieving outstanding results. This procedure is summarised in Fig. 5.

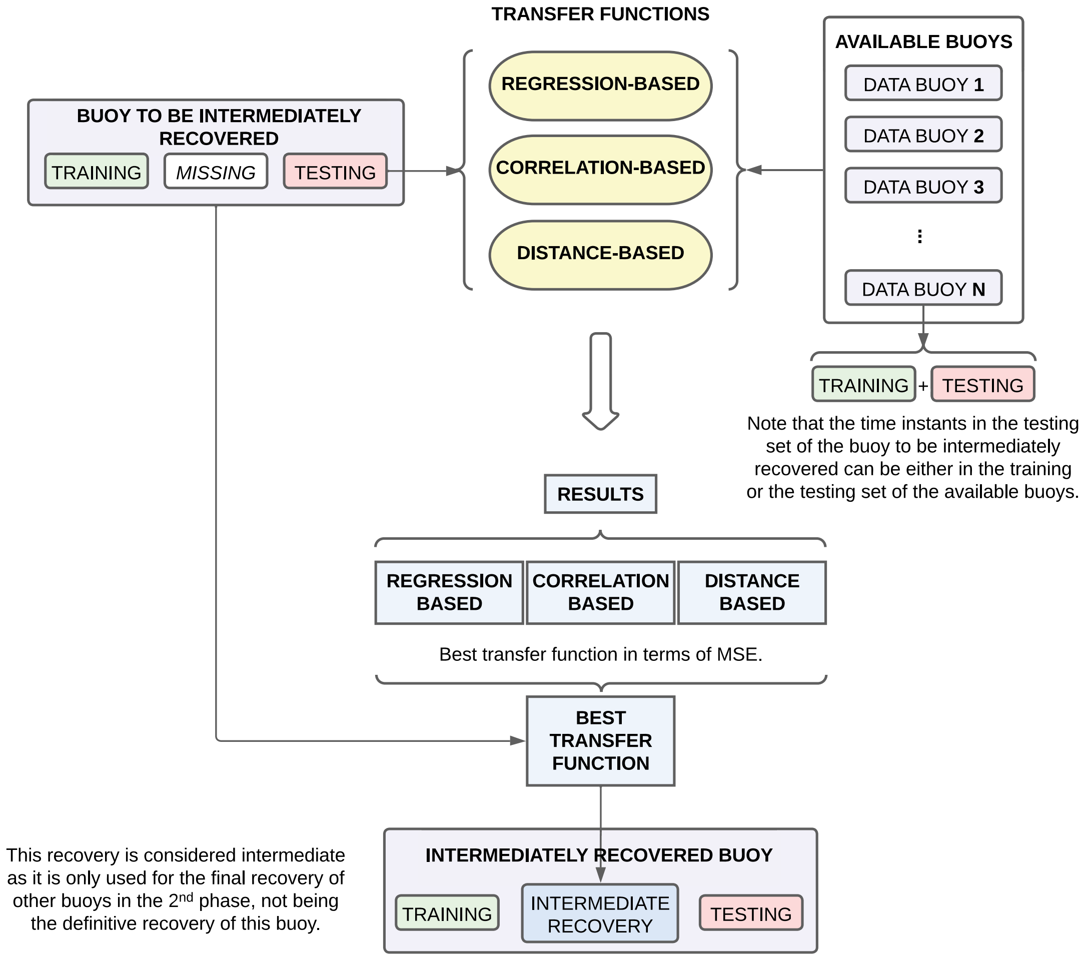
TRANSFER FUNCTIONS
REGRESSION-BASED
CORRELATION-BASED
DISTANCE-BASED
AVAILABLE BUOYS
DATA BUOY 1
DATA BUOY 2
DATA BUOY 3
⋮
DATA BUOY N
BUOY TO BE INTERMEDIATELY RECOVERED
TRAINING
MISSING
TESTING
TRAINING + TESTING
Note that the time instants in the testing set of the buoy to be intermediately recovered can be either in the training or the testing set of the available buoys.
RESULTS
REGRESSION BASED
CORRELATION BASED
DISTANCE BASED
Best transfer function in terms of MSE.
BEST TRANSFER FUNCTION
This recovery is considered intermediate as it is only used for the final recovery of other buoys in the 2nd phase, not being the definitive recovery of this buoy.
INTERMEDIATELY RECOVERED BUOY
TRAINING
INTERMEDIATE RECOVERY
TESTING

Fig. 4. Flowchart of the first phase in which buoys are individually recovered using one of the three TFs proposed. Once this phase finishes, the buoy is intermediately recovered. The term "intermediately" indicates that this reconstruction is not definitive and it is only used for recovering other buoys in the zone and not itself.

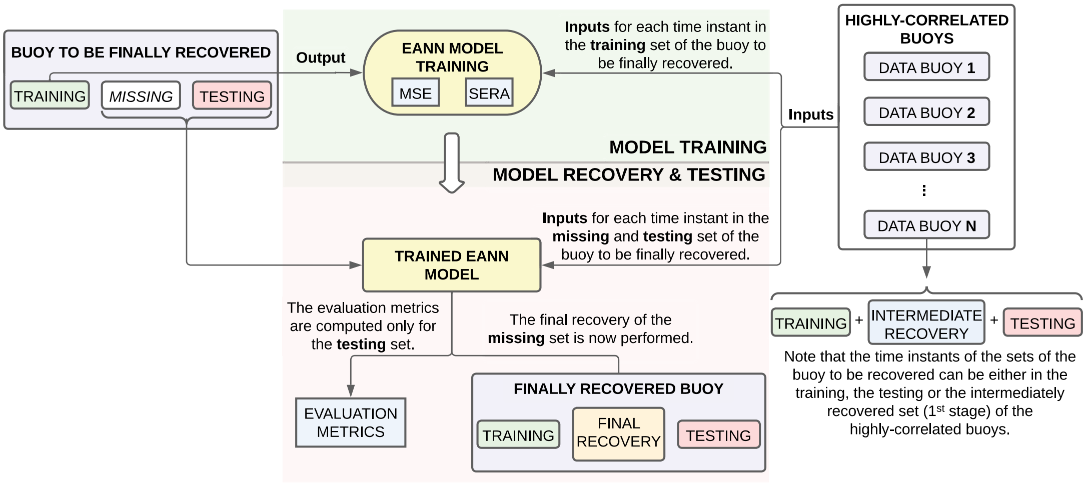
BUOY TO BE FINALLY RECOVERED
TRAINING
MISSING
TESTING
Output
EANN MODEL TRAINING
MSE
SERA
Inputs for each time instant in the training set of the buoy to be finally recovered.
Inputs
HIGHLY-CORRELATED BUOYS
DATA BUOY 1
DATA BUOY 2
DATA BUOY 3
⋮
DATA BUOY N
MODEL TRAINING
MODEL RECOVERY & TESTING
Inputs for each time instant in the missing and testing set of the buoy to be finally recovered.
TRAINED EANN MODEL
The evaluation metrics are computed only for the testing set.
The final recovery of the missing set is now performed.
EVALUATION METRICS
FINALLY RECOVERED BUOY
TRAINING
FINAL RECOVERY
TESTING
TRAINING + INTERMEDIATE RECOVERY + TESTING
Note that the time instants of the sets of the buoy to be recovered can be either in the training, the testing or the intermediately recovered set (1st stage) of the highly-correlated buoys.

Fig. 5. Flowchart of the second phase in which the final recovery of the buoy is carried out by means of the highly-correlated buoys of the zone. Once this phase finishes, the recovery of the buoy is considered definitive.

4.2.1. Artificial neural networks

Artificial Neural Networks (ANNs) are models which try to mimic the problem-solving behaviour of the human brain. Because of their powerful characteristics and properties, they are used in many real-world problems, being present in several applications of different areas of science. One of the simplest and most widely used models is Feed-Forward ANNs (FNNs) with a hidden layer composed of several nodes. An FFN is a generalisation of a

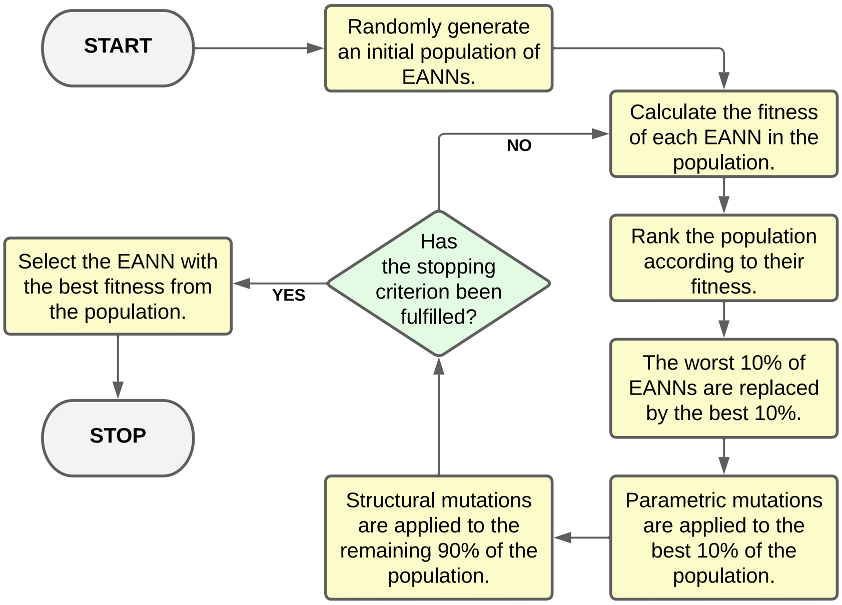
START
Randomly generate an initial population of EANNs.
Calculate the fitness of each EANN in the population.
Rank the population according to their fitness.
The worst 10% of EANNs are replaced by the best 10%.
Parametric mutations are applied to the best 10% of the population.
Structural mutations are applied to the remaining 90% of the population.
Has the stopping criterion been fulfilled?
NO
YES
Select the EANN with the best fitness from the population.
STOP

Fig. 6. Flowchart of the Evolutionary Algorithm applied for SWH time series reconstruction.

regression model where the basis functions are non-linear:

$$y(\mathbf{x}, \boldsymbol{\theta}) = \beta_0 + \sum_{d=1}^{D} \beta_d B_d(\mathbf{x}, \mathbf{w}_d), \tag{8}$$

where D is the number of hidden neurons, $\mathbf{w}_d$ represents the weights of the connections between the input layer and the hidden neuron d, $B_d(\mathbf{x}, \mathbf{w}_d)$ is the basis function of neuron d, which applies a non-linear transformation to the input space $\mathbf{x}$ (a vector containing the highly-correlated time series with respect to the one that is being finally recovered, i.e. those with a correlation above the threshold α), β_d is the weight of the connection between the d-hidden neuron and the output layer, β_0 represents the bias, and finally, the function to be optimised is denoted as $y(\mathbf{x}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{w})$.

Regarding the type of neurons in the hidden layer $B_d(\mathbf{x}, \mathbf{w}_d)$, Sigmoid Units (SUs), Product Units (PUs) and a combination of both of them (PU+SU) have been chosen in this work.

On the one hand, SUs basis functions [55] present an additive projection model. This family of units can approximate any given function with enough accuracy provided that the number of hidden neurons is selected appropriately. SU is one of the most used activation functions, and it has achieved excellent performance on a wide variety of problems. A SU is defined as:

$$B_d(\mathbf{x}, \mathbf{w}_d) = \frac{1}{1 + e^{-(w_{0,d} + \sum_{i=1}^{k} w_{i,d} x_i)}}, \tag{9}$$

where $w_{i,d}$ is the weight of the connection between the input i and the hidden neuron d, and $w_{0,d}$ is the bias.

On the other hand, PUs [56] are used for cases where there is a strong interaction between the inputs, and the decision regions are not separable into hyperplanes. PU based neural networks can form higher-order combinations of inputs, with the advantages of higher information capacity and smaller network architectures. PUs have been shown to work effectively on both classification and regression problems. A PU is formally defined as follows:

$$B_d(\mathbf{x}, \mathbf{w}_d) = \prod_{i=1}^{k} x_i^{w_{i,d}}. \tag{10}$$

In addition to develop ANNs using each type of basis function, this work also proposes a hybridisation in which the hidden layer is made up of neurons of both types, PUs and SUs (known as PU+SU), trying to benefit from the advantages of both.

4.2.2. Evolutionary artificial neural networks

Back Propagation (BP) algorithm is the most widespread one to train ANNs. Nonetheless, this algorithm only optimises the connection weights given a predefined neural network structure. In addition, it is challenging to define the most suitable structure for each problem in advance. Besides, BP can converge to local optima due to the convoluted error surface associated with ANNs. Therefore, in this work, an Evolutionary Algorithm (EA) is used to train both the structure and the connection weights of ANNs, giving rise to Evolutionary ANNs (EANNs). The goal of EANNs is to perform a balanced search of the error surface (i.e. exploration vs exploitation) to avoid local minima and find good performance solutions. For this purpose, mutation operators are used aiming to increase the diversity and exploitation of EANNs during their evolution. Given that the EA is not gradient-based, parameters such as learning rate and momentum are not used. Instead, the control parameters of the mutation operators regulate the strength when altering the synaptic weights and the structure of the EANNs. Since crossover operator has been proven to lead to potential drawbacks in evolving EANNs [57], it is not considered in the EA.

Specifically, the EA used in this work evolves the ANNs proposed in Section 4.2.1 by applying the evolutionary process, which is graphically summarised in Fig. 6. As can be seen, the EA starts creating an initial random population of EANNs. In this way, the structure of each EANN in the population is randomly created, that is, its number of hidden neurons, the number of connections linking each hidden neuron to both input layer and output layer, and the synaptic weight of each connection are randomly initialised according to the values of the parameters shown in Table 3.

After that, the evolutionary process is performed on EANNs, generation after generation, until the stopping criterion is reached. The performance of EANNs is evaluated in each generation of the evolutionary process by calculating their fitness (Eq. (11)), which is used to sort the population (the higher fitness, the better). Once EANNs are sorted, the worst 10% of the population is replaced by a copy of the best 10% of EANNs, and then each EANN in the population is individually evolved. Specifically,

Table 3
Most important parameters values considered of the EA for MSE and SERA.

Parameter	PU	SU	PU+SU
Maximum number of generations.[a]	1200/2100	3800/1800	1600/2100
Population size of ANNs.	1000	1000	1000
Minimum number of hidden neurons.	1	1	1
Maximum number of hidden neurons.[a]	4/6	6/6	5/6
Synaptic weights for connections between input and hidden layer.	$[-5, 5]$	$[-10, 10]$	$[-10, 10]$
Synaptic weights for connections between hidden and output layer.	$[-10, 10]$	$[-10, 10]$	$[-10, 10]$
Number of hidden neurons to add or remove.	$[1, 2]$	$[1, 2]$	$[1, 2]$
Percentage of hidden layer connections to add or remove.	30%	30%	30%
Percentage of output layer connections to add or remove.	5%	5%	5%

[a] Optimising by MSE/SERA.

the best 10% of the population is parametrically mutated (altering the synaptic weights), whereas the remaining 90% is structurally mutated (adding or removing hidden neurons and connections). It is noteworthy that the mentioned selection of the best 10% of EANNs that replace the worst 10% of the population encourages elitism in the EA since after that replacement, the worst 90% of EANNs contain the best 10% of EANNs and, as a consequence, the best 10% of EANNs are independently evolved in both ways: parametrically and structurally. Parametric and structural mutations will be later described.

Thus, each EANN in the population is optimised by maximising its fitness (i.e. improving its performance) during the evolutionary process. Finally, after reaching the stopping criterion, the EA stops and returns the best EANN according to the fitness.

Given that the goal of the EA is to maximise the fitness of each EANN in the population throughout the evolution, the fitness function used by the EA is expressed as follows:

$$F(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{1 + \Phi(\mathbf{x}, \boldsymbol{\theta})}, \tag{11}$$

where Φ is the metric used to compute the error achieved by the EANN. Specifically, to optimise the fitness of the EANNs two metrics are considered, the Mean Squared Error (MSE) and the Squared Error-Relevance Area (SERA), which are described in Section 4.2.3.

Therefore, taking into account that the fitness of each EANN is computed as a decreasing transformation of its error (as expressed in Eq. (11)), the best fitness EANN returned by the EA corresponds to the SWH reconstructed time series with the lowest estimation error.

As mentioned above, the purpose of the mutation operators is to increase the diversity and exploitation of the population. To this end, each EANN is individually mutated and the strength of each mutation during its evolution depends on its associated temperature, defined by:

$$T(\mathbf{x}, \boldsymbol{\theta}) = 1 - F(\mathbf{x}, \boldsymbol{\theta}), \qquad 0 \leq T(\mathbf{x}, \boldsymbol{\theta}) < 1. \tag{12}$$

On the one hand, the parametric mutation involves altering the synaptic weights of the EANNs (not the structure), and it is performed on the best 10% of the population. It is carried out by adding Gaussian noise of zero mean and decreasing variance throughout the evolution depending on the temperature of each EANN so that the EA moves from exploring solutions to exploiting them. In that way, the synaptic weights of the connections from the input layer to the hidden layer are modified as follows [58]:

$$w_{i,d}(t+1) = w_{i,d}(t) + \xi_1(t), \qquad i = 1, \dots, k, d = 1, \dots, D, \tag{13}$$

where $\xi_1(t) \in N(0, \ \alpha_1(t) \cdot T(\mathbf{x}, \boldsymbol{\theta}))$ corresponds to a number randomly generated according to a normal distribution of one dimension, whose mean and variance are equal to 0 and $\alpha_1(t) \cdot T(\mathbf{x}, \boldsymbol{\theta})$, respectively. The goal is that the strength of the mutations lessens as the EANNs increase their fitness, but also managed by an adaptive parameter $\alpha_1(t)$ that will be later described.

The synaptic weights of the connections from the hidden layer to the output layer are modified as follows [58,59]:

$$\beta_d(t+1) = \beta_d(t) + \xi_2(t), \qquad d = 1, \dots, D, \tag{14}$$

being $\xi_2(t) \in N(0, \ \alpha_2(t) \cdot T(\mathbf{x}, \boldsymbol{\theta}))$ similar to $\xi_1(t)$ but, for this case, a different control parameter is considered for the variance $\alpha_2(t)$.

After applying the parametric mutation, the EANN fitness is recalculated, and then the mutation is rejected or accepted according to a simulated annealing process [60]. Specifically, being ΔF, the difference of the EANN fitness before and after the mutation, the mutation is accepted if $\Delta F \geq 0$. On the contrary, if the new fitness is worse than the original, the mutation would be accepted with a probability $\exp(\Delta F / T(\mathbf{x}, \boldsymbol{\theta}))$.

The two aforementioned adaptive parameters $\alpha_1(t)$ and $\alpha_2(t)$ control the strength of parametric mutations during the evolution of EANNs. To proceed with, both parameters are updated during evolution to avoid local minima and accelerate convergence when search conditions are suitable. The update of $\alpha_1(t)$ and $\alpha_2(t)$ is expressed as follows [56]:

$$\alpha_k(t) = \begin{cases} (1+\lambda) \cdot \alpha_k(t) \\ \quad \text{if} \quad F(\mathbf{x}, \boldsymbol{\theta}_g) > F(\mathbf{x}, \boldsymbol{\theta}_{g-1}) \ \forall g \in \{t, t-1, \dots, t-\rho\} \\ (1-\lambda) \cdot \alpha_k(t) \\ \quad \text{if} \quad F(\mathbf{x}, \boldsymbol{\theta}_g) = F(\mathbf{x}, \boldsymbol{\theta}_{g-1}) \ \forall g \in \{t, t-1, \dots, t-\rho\} \\ \ \alpha_k(t) \ \text{otherwise} \end{cases}, \tag{15}$$

where $k \in \{1, 2\}$, $F(\mathbf{x}, \boldsymbol{\theta}_g)$ is the best EANN fitness in generation g, and λ and ρ are parameters to control the update. Their values are set to $\alpha_1(0) = 0.5$, $\alpha_2(0) = 1$, $\lambda = 0.1$, and $\rho = 10$. The use of Eq. (15) is justified as follows: a successful generation means that the current best EANN is better than one of the previous generation. When this occurs ρ times, the best EANNs are most likely to be found in the search space being explored. Consequently, the strength of the mutation is increased with the aim of finding EANNs closer to the optimal one. Conversely, the mutation strength decreases when the best EANN is the same during ρ generations. Otherwise, the strength of the mutation remains the same.

On the other hand, the structural mutation involves altering the structure of the EANNs (adding or removing both hidden neurons and the connections that link them to the input layer and output layer), and it is performed on the remaining 90% of the population. In that way, the EA explores a diverse range of structures (expands the area of the search space) and keeps a diverse population.

Five different structural mutations are used by the EA: Add neuron, Delete neuron, Add connection, Delete connection and Neuron fusion, which are sequentially performed on each EANN with probability $T(\mathbf{x}, \boldsymbol{\theta})$. In case no structural mutation is performed due to probability, one of them is randomly selected and then performed. After that, the fitness of the EANN is recalculated.

The connection structural mutations are performed as follows [58,59]:

- Add connection. Two neurons from adjacent layers are randomly selected, and then both neurons are connected with a random synaptic weight. This mutation is firstly performed to add a connection between two neurons of input and hidden layers, and then another connection is added to link two neurons of hidden and output layers.
- Delete connection. This mutation is also applied to all adjacent layers, randomly selecting one neuron belonging to the mutated layer and another neuron from the previous layer.

The number of connections used for both mutations is obtained as $\Delta_{min} + u \cdot T(\mathbf{x}, \boldsymbol{\theta}) \cdot [\Delta_{max} - \Delta_{min}]$, being u a number randomly generated between 0 and 1, $\Delta_{min} = 1$ is the minimum number of connections to add or delete, and Δ_{max} is the maximum number of connections to mutate, which is calculated by multiplying a user-defined percentage by the total number of connections of the layer being mutated.

Finally, the neuron structural mutations are performed as follows [60]:

- Neuron addition. One or more hidden neurons are added. Two neurons are randomly selected, one from the input layer and one from the output layer. Then, both neurons and the new one added in the hidden layer are connected with random synaptic weights. Specifically, the synaptic weights for the connections from the input layer to the hidden layer are selected in the range $[-I, I]$, whereas for the connections from the hidden layer to the output layer, the range $[-O, O]$ is used. Both ranges values are user-defined.
- Neuron deletion. One or more hidden neurons are randomly selected and removed along with their connections.
- Neuron fusion. Two neurons, a and b, are randomly chosen and fused into one neuron c, keeping their common connections and whose synaptic weights are recalculated as follows [60]:

  $$\beta_c = \beta_a + \beta_b, w_{i,c} = \frac{w_{i,a} + w_{i,b}}{2}. \tag{16}$$

  Non-common connections between neurons a and b are inherited by c with a probability of 0.5, keeping their original values.

The number of neurons used for these mutations is obtained as $\Delta_{min} + u \cdot T(\mathbf{x}, \boldsymbol{\theta}) \cdot [\Delta_{max} - \Delta_{min}]$, where Δ_{min} and Δ_{max} are user-defined and represent the minimum and maximum number of neurons to be mutated, respectively.

Finally, if the mutated EANN is not valid, all applied mutations are discarded, and another structural or parametric mutation is randomly chosen and performed on the original EANN, avoiding the use of repair mechanisms.

4.2.3. Metrics

For regression problems, one of the most commonly used metrics is the Mean Squared Error (MSE), which is computed as follows:

$$MSE(\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}(Y_t - \hat{Y}_t)^2, \tag{17}$$

where T is the number of time instants, Y_t is the observed SWH value at t time instant, and $\hat{Y}_t$ is the recovered SWH value by the model ($\hat{Y}_t = y(\mathbf{x}, \boldsymbol{\theta})$).

Although it is true that this metric is well-known, it has a severe drawback: as it computes a mean value of the errors, very high or extreme values of time series tend to be underestimated. However, note that these very high or extreme values are the most important ones in this type of time series, as not correctly predicting them may entail for example, not anticipating seaquakes, among other situations. Therefore, the recovery of these values is of enormous interest for the marine engineering. In this sense, and given that not all values of the target variable are equally important, this work proposes using the metric Squared Error-Relevance Area (SERA) [27] to overcome this drawback.

The following calculations are considered to obtain this metric. Firstly, it is important to mention that the interval proposed by [61] frames the mean values. Values outside this interval are considered outliers. In our case, these are the extreme SWHs. Specifically, the interval for determining the outlier cut-off values is $[Q1 - 1.5IQR,\ Q_3 + 1.5IQR]$, where Q_1 and Q_3 are the first and third quartiles of the SWH time series Y, and $IQR = Q_3 - Q_1$ is the interquartile range. Note that, in this work, the interest is in the right part of the interval, that is, to focus on extreme values higher than $Q_3 + 1.5IQR$. To make this interval less proned to bias, in [62], the medcouple metric (MC) [63] was included as a robust alternative to the classical skewness coefficient:

$$MC = med_{Y_i \leq Q_2 \leq Y_j} h(Y_i, Y_j), \tag{18}$$

where Q_2 is the median of the SWH time series Y, and for all $Y_i \neq Y_j$, h is computed as:

$$h(Y_i, Y_j) = \frac{(Y_j - Q_2) - (Q_2 - Y_i)}{Y_j - Y_i}. \tag{19}$$

After that, to determine the cut-off values for the outliers, in [62], the authors proposed an interval, where a threshold γ is obtained, defined as:

$$\gamma = \begin{cases} Q_3 + 1.5e^{3MC}IQR & \text{if} \quad MC \geq 0 \\ Q_3 + 1.5e^{4MC}IQR & \text{if} \quad MC < 0 \end{cases}. \tag{20}$$

According to [62], these exponential functions allow the boxplot to be more adjusted to the skewness. In the context of imbalanced regression tasks, they present two main contributions: (1) the metric is non-parametric, and (2) the method is better suited to avoid missing real cases of extreme values.

Then, a relevance value r_t is assigned to each value Y_t, according to the following rules:

$$r_t = \begin{cases} 0 & \text{if} \quad Y_t \leq MC \\ 1 & \text{if} \quad Y_t \geq \gamma \\ \frac{Y_t - Q_2}{\gamma - Q_2} & \text{if} \quad Q_2 \leq Y_t \leq \gamma \end{cases}. \tag{21}$$

Considering the subset S formed by the cases for which the relevance value r_t assigned to the target value Y_t is above or equal to a cut-off s, the Squared Error-Relevance (SER) of a model with respect to this cut-off s is formulated as:

$$SER_s(\mathbf{x}, \boldsymbol{\theta}) = \sum_{t \in S}(\hat{Y}_t - Y_t)^2. \tag{22}$$

Finally, SERA represents the area below the SER_s:

$$SERA(\mathbf{x}, \boldsymbol{\theta}) = \int_0^1 SER_s(\mathbf{x}, \boldsymbol{\theta})ds = \int_0^1 \sum_{t \in S}(\hat{Y}_t - Y_t)^2 ds. \tag{23}$$

Note that the smaller is the area under this curve, the better the model is.

According to the study carried out in [27], using the Mean Squared Error (MSE) to compare two models can be problematic. If one model performs better around the mean and the other is better at extreme values, the MSE may not show any difference between them. The reason behind this behaviour is that MSE do not appropriate reflect the difficulty in correctly estimating an extreme value in comparison to predicting a common value around mean. This happens because the MSE considers all values equally important, even though predicting an extreme value is considerably more difficult. On the other hand, if we only focus on extreme values, we ignore the performance around the mean. To address this issue, SERA was introduced as a solution. SERA orders values according to their relevance. A higher relevance is given to extreme values than to more common values. For this, we use a threshold value s. The SER$_s$ metric is then used to calculate the error only for values whose relevance is above s. Combining SER$_s$ and MSE is the suggested solution by SERA to avoid neglecting the impact of model performance on any of the metrics. The errors of the around-mean cases have fewer impact given their higher frequency. On the contrary, the errors made when estimating extreme values are given more importance and are counted more times along the cut-off values of relevance in the global sum defined in the integral.

Moreover, the SERA metric demonstrates a notable advantage in handling severe model biasing, as it replaces the conventional skewness coefficient with the medcouple, thereby enhancing its ability to detect and account for extreme values (outliers). Furthermore, SERA exhibits all the essential attributes necessary for an appropriate metric in the context of imbalanced regression problems: (1) it primarily focuses on reducing prediction errors associated with extreme target values; (2) it mitigates the risk of overfitting by giving equal importance on predictions for common values; (3) it enables the computation of asymmetric errors, i.e. errors of similar magnitude have different impact according to their relevance; and (4) SERA facilitates model discrimination and comparison. Therefore, given its inherent properties, the SERA metric is regarded as a suitable and effective method for evaluating imbalanced regression problems [27].

4.3. Complexity of the proposed methodology

In order to determine the complexity of the proposed method O_{method}, we must consider the complexities of the two phases discussed earlier.

Let us denote as O_{phase1}, the complexity of the first phase. This first phase sequentially applies three transfer function models. Considering N as the number of time series involved in the method: the linear regression-based model is an $O(N^2)$ algorithm thanks to advanced matrix inversion algorithms; the complexity of the correlation-based model is $O(N^2)$ since for each time series the correlation with all the other time series should be calculated; and the distance-based model requires to compute the distance between each pair of time series so the complexity is $O(N^2)$. Thus, applying the properties of Big O notation, $O_{phase1}(N^2 + N^2 + N^2) \simeq O_{phase1}(N^2)$.

The computational cost of the second phase, related to the evolutionary algorithm, is $O_{phase2}(gPc)$, where g is the number of generations, P is the population size, and c is the size of the individual. This order of complexity is obtained since each generation of the evolutionary algorithm involves evaluating and ranking the population and updating it for the next generation, through structural and parametric mutations.

In the worst case, assuming that $n = \max(N, g, P, c)$, the complexity $O_{method}(n^2 + n^3) \simeq O_{method}(n^3)$. Therefore, it can be concluded that the methodology is of cubic order.

5. Experimental settings and results

This Section describes the experimental settings used in both phases of the proposed methodology. Besides, the results obtained for both coastal areas and a detailed discussion comparing the methodology with other state-of-the-art algorithms are also presented.

5.1. Experimental settings

As mentioned in Section 4, the proposed methodology is divided into two phases. The first performs an intermediate recovery of each SWH time series, serving as input for the final recovery of other buoys in the zone.

On the one hand, for the first phase, three different TFs are proposed, which have been optimised by MSE. Regarding the regression-based TF, the initial correlation threshold α is determined using a 10-fold cross-validation over the training sets and using the following grid $\{0.65, 0.70, 0.75\}$, which, in turn, is selected because these values are close to the mean of the correlations between the training sets of the buoys (Tables 5 and 6). Note that in each step of the regression-based TF, this threshold α is increased by 0.05 to reduce the error entailed in the reconstruction process. In this sense, the initial values $\alpha = 0.70$ and $\alpha = 0.65$ are chosen for the Gulf of Alaska and the Northeast Coast zones, respectively. The other two TFs (correlation-based and distance-based) do not need any parameter to be cross-validated, just the correlations and distances between the buoys belonging to the same coastal zone, which are precomputed and shown in Tables 5 and 6.

As it was mentioned at the end of Section 4.1, it may be assumed that when correlation increases, distance decreases and vice-versa. Looking at Tables 5 and 6, it can be observed that for the buoy 44025, the highest correlation (0.938) is found with the buoy 44065, which is also the closest (47.549 km). In addition, the second highest correlation (0.905) is found with the buoy 44066, which is also the second closest buoy (83.186 km). Thus, the general assumption is met in some cases. However, some cases deny this assumption in two ways. (1) The higher the correlation, the lower the distance: in this case, highly correlated buoys are supposed to be geographically close to each other. However, looking at the three buoys located at the coast of the Gulf of Alaska (46076, 46061, 46082, Table 5), it can be seen that the buoy 46076, which is geographically located at the left side of the coast, has a higher correlation with the buoy in the opposite extreme (buoy 46082, correlation of 0.833), than with respect to the buoy 46061, which is geographically located at the middle of both (0.825). The same behaviour is replicated with buoys 44020, 44013, and 44066 of the Northeast Coast (Table 6). The same correlation is obtained for the pairs 44020 − 44013 (0.704) and 44020 − 44066 (0.703), whereas the distance between the buoys of the first pair (99.718 km) is almost three times the distance between the buoys of the second pair (288.749 km). (2) The higher the distance, the lower the correlation: this is the opposed situation, which is also generally assumed. This idea may not be deduced either as distance does not take into account the geographical features that may exist or the orography of the coast. For instance, focusing on the buoy 44005 of the Northeast Coast, it can be seen that its closest buoy is the 44013 (156.558 km away). However, that buoy is not the most correlated (0.782, in fact, is the third). The same behaviour could be found for some of the remaining buoys.

On the other hand, for the second phase, the EA used to optimise the ANN models has been applied using each optimisation metric (MSE and SERA) independently. More specifically, Table 3 shows the most important parameters considered and their range of values according to the basis functions used.

Since the EA is stochastic, 30 runs for each metric (MSE and SERA), have been carried out, using different seeds, to recover each SWH time series. It is worthy of mention that the number

Table 4
Considered range of values for tuning the parameters of the state-of-the-art techniques.

Technique	Parameter description	Range of values
Ridge	Regularisation	$\{10^{-3}, 10^{-2}, \dots, 10^{3}\}$
Lasso	Regularisation	$\{10^{-3}, 10^{-2}, \dots, 10^{3}\}$
ElasticNet	Regularisation	$\{10^{-3}, 10^{-2}, \dots, 10^{3}\}$
	Ratio of the L1 penalisation weight	[0.10, 0.50, 0.70, 0.90, 0.95, 0.99, 1.00]
SVR	Kernel width	$\{10^{-3}, 10^{-2}, \dots, 10^{3}\}$
MLP	Number of hidden neurons	[10, 25, 50, 100]
	Regularisation	$\{10^{-3}, 10^{-2}, \dots, 10^{3}\}$
	Initial learning rate	[0.00001, 0.0001, 0.001]
	Number of iterations	[1000, 1500]
RandomForest	Number of trees in the forest	[10, 50, 100, 150]
	Maximum depth of the trees	[3, 4, 5, 6]
	Maximum depth of the trees: when all leaves are pure or contain less than samples than the minimum	[2, 3, 4, 5]

of generations required for the evolutionary process depends on the time series being recovered.

Furthermore, in the second phase, with the aim of establishing a fair and robust comparison of the EANN models, 7 state-of-the-art techniques have been run: LinearRegression [64], Ridge [65], Lasso [66], ElasticNet [67], Support Vector Machines (SVR) [68], Multilayer Perceptron (MLP) [69] and RandomForest [70]. Their parameters values have been selected using a 10-fold cross-validation over the training sets. The best configuration for each technique is chosen according to the metric being optimised (i.e. lowest MSE or SERA). More specifically, Table 4 shows the most important parameters considered and their values. Both MLP and RandomForest techniques are stochastic, and thus, they have been run 30 times using different seeds as the EANN models. Therefore, their results are expressed as their mean and Standard Deviation (*SD*): $Mean_{SD}$.

Finally, The performances of the TFs in the first phase and the ML techniques in the second phase are evaluated considering three metrics in addition to the two used to optimise the models (i.e. MSE and SERA, both to be minimised). These three new metrics are the standard coefficient of determination (R^2), the Maximum Standard Error (MaxSE) and the Mean Relative Error (MRE). The first one, R^2, is a measure of goodness of fit between 0 and 1 that indicates how well a model predicts an outcome. The second one, MaxSE, indicates the biggest error performed by the model, hence, the lower error, the better. It is useful to know the magnitude of the errors made by the models. It is computed as follows:

$$MaxSE(\mathbf{x}, \boldsymbol{\theta}) = \max(Y_t - \hat{Y}_t)^2, \quad t \in \{1, 2, \dots, T\}, \tag{24}$$

where T is the number of time instants, Y_t is the observed SWH value at t time instant and $\hat{Y}_t$ is the recovered SWH value by the model ($\hat{Y}_t = y(\mathbf{x}, \boldsymbol{\theta})$).

Finally, MRE indicates the ratio of the absolute error with respect to the magnitude of the measurement being taken. MRE is calculated as follows:

$$MRE(\mathbf{x}, \boldsymbol{\theta}) = \frac{\sum_{t=1}^{T} |Y_t - \hat{Y}_t|}{\sum_{t=1}^{T} |Y_t|}. \tag{25}$$

These last two metrics, MaxSE and MRE, allow the comparison of the results obtained for different buoys. Even though SWH is always measured using the same unit, the magnitudes from one buoy to others differ depending on the atmospheric and geographical conditions where they are located.

5.2. Results

As the proposed methodology is divided into two different phases, the results obtained using the test sets are presented separately for each phase. Concretely, the results achieved by the three TF models proposed for the first phase are shown in Table 7. As can be seen, the results are divided according to the zone under study, i.e. the first part of the Table 7 shows the results for the Gulf of Alaska, whereas the results for the Northeast Coast are shown below. In addition to MSE, which has been used to optimise the models, the performance of each model in terms of SERA, R^2, MaxSE, and MRE is also shown for comparison purposes, all of them to be minimised but R^2. Regarding the Gulf of Alaska, even though the correlation-based TF achieves good results (the number of second-best results can denote it), the distance-based TF manages to improve these results. For this zone, the regression technique is selected for 3 out of 5 buoys, whereas the distance one is selected for the remaining 2 buoys. In the case of the Northeast Coast, all these TF models can achieve competitive results for all the performance measures. Focusing on MSE, the regression-based TF is chosen 1 time, the correlation-based TF is chosen 3 times, and the distance-based is selected 4 times. Ultimately, for all the buoys of this work, the regression-based, the correlation-based and the distance-based TFs are chosen a total of 4, 3, and 6 times, respectively. This demonstrates that weighting by distance between buoys, which is one of the contributions of this study, is better than weighting by their similarity. Note that, as mentioned in Section 4.1, distance and correlation (similarity) may not be related in such a way that the greater distance, the lowest correlation or vice-versa. Not assuming this fact has led to achieving better results in 6 buoys. Moreover, it is important to mention that MRE supports the results achieved by MSE so that the selected TF for each buoy would be the same but for one case (44008). Therefore, the results achieved in terms of MSE and MRE are consistent.

Therefore, the intermediate recovery of each of the analysed buoys has been selected according to the TF, achieving the best result in terms of MSE. For instance, for the buoy 44005 of the Northeast Coast, the distance-based TF achieves the best result (0.2242), improving the performance, in terms of MSE, of the remaining TFs (0.3945 and 0.2431 for the regression-based and the correlation-based TFs, respectively). As aforementioned, these intermediate reconstructions are not definitive and are only used for the final reconstruction of the other buoys belonging to the same coastal zone but not for themselves.

In this way, Table 8 shows the results obtained by the techniques applied in the second phase when optimised by MSE, whereas Table 9 contains the results when SERA optimises the techniques. Note that the results are expressed as their mean and Standard Deviation (*SD*): $Mean_{SD}$ for the stochastic techniques. The first part of each table shows the results for the Gulf of Alaska, whereas the results for the Northeast Coast are shown below. Our approach consists in the EANN models, which are named

Table 5
Correlation and distance matrices for the buoys of the Gulf of Alaska.

	Buoy	46001	46061	46076	46078	46082	46085
Correlation matrix	46001	1	0.674	0.719	0.804	0.682	0.833
	46061	0.674	1	0.825	0.591	0.872	0.666
	46076	0.719	0.825	1	0.625	0.833	0.679
	46078	0.804	0.591	0.625	1	0.515	0.626
	46082	0.682	0.872	0.833	0.515	1	0.749
	46085	0.833	0.666	0.679	0.626	0.749	1
	Buoy	46001	46061	46076	46078	46082	46085
Distance matrix	46001	0	450.195	360.178	298.422	468.802	316.905
	46061	450.195	0	107.636	621.154	202.341	536.914
	46076	360.178	107.636	0	513.643	262.091	501.788
	46078	298.422	621.154	513.643	0	713.890	608.097
	46082	468.802	202.341	262.091	713.890	0	423.313
	46085	316.905	536.914	501.788	608.097	423.313	0

Distance values are expressed in kms.

Table 6
Correlation and distance matrices for the buoys of the Northeast Coast.

	Buoy	44005	44008	44011	44013	44020	44025	44027	44065	44066
Correlation matrix	44005	1	0.811	0.766	0.782	0.679	0.631	0.840	0.534	0.719
	44008	0.811	1	0.829	0.745	0.726	0.726	0.694	0.618	0.824
	44011	0.766	0.829	1	0.626	0.574	0.509	0.652	0.404	0.614
	44013	0.782	0.745	0.626	1	0.704	0.626	0.480	0.616	0.693
	44020	0.679	0.726	0.574	0.704	1	0.674	0.488	0.620	0.703
	44025	0.631	0.726	0.509	0.626	0.674	1	0.496	0.938	0.905
	44027	0.840	0.694	0.652	0.480	0.488	0.496	1	0.379	0.555
	44065	0.534	0.618	0.404	0.616	0.620	0.938	0.379	1	0.826
	44066	0.719	0.824	0.614	0.693	0.703	0.905	0.555	0.826	1
	Buoy	44005	44008	44011	44013	44020	44025	44027	44065	44066
Distance matrix (km)	44005	0	300.735	315.658	156.558	212.204	468.762	189.772	492.931	494.625
	44008	300.735	0	235.818	236.321	140.299	332.583	450.303	377.037	304.874
	44011	315.658	235.818	0	366.812	313.683	564.459	359.799	606.895	540.689
	44013	156.558	236.321	366.812	0	99.718	313.560	346.205	336.430	346.389
	44020	212.204	140.299	313.683	99.718	0	279.113	393.829	313.592	288.749
	44025	468.762	332.583	564.459	313.560	279.113	0	658.341	47.549	83.186
	44027	189.772	450.303	359.799	346.205	393.829	658.341	0	682.634	681.140
	44065	492.931	377.037	606.895	336.430	313.592	47.549	682.634	0	122.930
	44066	494.625	304.874	540.689	346.389	288.749	83.186	681.140	122.930	0

Distance values are expressed in kms.

according to the basis functions used in the hidden layer: PU for Product Units, SU for Sigmoid Units, and PU+SU for hybrid EANNs with Product Units and Sigmoid Units. As an example, the buoy 46085 of the Gulf of Alaska uses the buoys 46001 and 46082 as inputs for its final recovery, given that the correlations between the buoy to be finally recovered (46085) and these two buoys (0.833 and 0.749, respectively, as can be checked from Table 5) are the only ones that are above 0.70, which is the threshold α for this zone.

On the one hand, analysing the results obtained when optimising the models by MSE (Table 8), it can be observed that in the Gulf of Alaska, EANNs models obtained the best results in 4 out of 5 buoys (2 using SU and 3 using PU+SU, note that SU and PU+SU share one best result), whereas the remaining best result is achieved by both LinearRegression and Ridge (buoy 46078). Regarding the Northeast Coast, EANN models are able to achieve the best MSE results in 5 out of 8 buoys (1 using PU, 3 using SU and 2 using PU+SU, note that SU and PU+SU share one best result), whereas the remaining 3 best results are achieved by RandomForest (buoy 44011), SVR (buoy 44027) and regression techniques (the best result for the buoy 44066 is achieved by both LinearRegression and Ridge). Note that EANN models can also obtain the second-best results in 2 out of those 3 buoys. The results achieved in this second phase are better than those obtained in the first one for all the 13 buoys (5 for the Gulf of Alaska and 8 for the Northeast Coast).

Moreover, when models are optimised by MSE, EANN models obtained not only the best SERA results in both areas (5 for PU, 1 for SU and 7 for PU+SU) but also the second best results. These results in terms of SERA are very interesting since this metric involves recovering not only the simpler parts where the SWH time series takes values around the mean but also the more complex parts where the SWH time series takes extreme values. In this sense, it can be said that EANNs using PUs in the hidden layer can capture both behaviours of the SWH time series on the simplest buoys (i.e. buoys with a reduced percentage of missing values or buoys with few extreme values), and when buoys are considered to be more difficult to retrieve (i.e. buoys with a high percentage of missing values or buoys with a lot of extreme values), PUs combined with SUs, i.e. hybrid PU+SU models, are able to achieve the best results on most buoys. In terms of R^2, the results are in line with those obtained by MSE. Interestingly, most of the buoys reach R^2 values around 0.8 and some buoys above 0.9, except for two buoys of the Northeast Coast (44011 and 44027), for which values of 0.5265 and 0.3574, respectively, have been reached. The reason behind these poor values obtained by buoys 44011 and 44027 is that the correlation with which the input buoys are selected is very close to the correlation threshold chosen for this zone, $\alpha = 0.65$, and only 3 and 1 buoys are used in the final recovery, respectively, as can be checked in Table 6. Analysing the MaxSE results, it can also be observed that the highest values of SE (Squared Errors) are obtained for the same two buoys at the Northeast Coast (44011 and 44027) and for

Table 7
Results achieved by the three TF models proposed for the first phase of the approach, optimising models by MSE.

	Metric	Buoy	TF Model		
			Regression	Correlation	Distance
Gulf of Alaska	MSE (↓)	46061	**0.1674**	0.2385	*0.1833*
		46076	0.5994	*0.2852*	**0.1925**
		46078	**0.4452**	0.6282	*0.5735*
		46082	0.8910	*0.4649*	**0.4026**
		46085	**0.5579**	0.6426	*0.5989*
	SERA (↓)	46061	**11502.2972**	18638.2711	*14393.3618*
		46076	56501.5219	*21347.8961*	**15012.6605**
		46078	**29039.6746**	41748.0934	*37639.7270*
		46082	51038.8417	*20229.0578*	**17466.1147**
		46085	**33107.1304**	39644.6752	*36518.7132*
	R^2 (↑)	46061	**0.8465**	0.7956	*0.8424*
		46076	0.5537	*0.7947*	**0.8612**
		46078	**0.7194**	*0.6332*	0.5990
		46082	0.4470	*0.6963*	**0.7415**
		46085	*0.7366*	0.7229	**0.7374**
	MaxSE (↓)	46061	**6.4872**	9.4614	*8.0193*
		46076	25.1624	*14.9769*	**10.8107**
		46078	**15.6620**	22.3124	*20.8965*
		46082	22.5341	*8.4985*	**7.1036**
		46085	12.9483	*9.8247*	**9.3184**
	MRE (↓)	46061	**0.1746**	0.2080	*0.1790*
		46076	0.2929	*0.2089*	**0.1694**
		46078	**0.2034**	0.2372	*0.2262*
		46082	0.2719	*0.2020*	**0.1870**
		46085	**0.1841**	0.1977	*0.1913*
	Metric	Buoy	TF Model		
			Regression	Correlation	Distance
Northeast Coast	MSE (↓)	44005	0.3945	*0.2431*	**0.2242**
		44008	0.3519	**0.2043**	*0.2070*
		44011	1.1363	**1.0015**	*1.0434*
		44020	0.0531	**0.0443**	*0.0452*
		44025	0.1677	*0.1360*	**0.0491**
		44027	**0.4869**	*0.5077*	0.5223
		44065	*0.1063*	0.1283	**0.0682**
		44066	0.5848	*0.2280*	**0.1608**
	SERA (↓)	44005	25798.3631	*14717.2076*	**13460.3862**
		44008	13360.9987	**9464.4901**	*9901.0025*
		44011	49114.9313	**33628.0406**	*34449.8383*
		44020	4749.2202	*3390.8779*	**3347.2782**
		44025	11325.0573	*10216.2087*	**3296.0708**
		44027	46884.9705	**38049.4116**	*40616.0840*
		44065	*8469.1762*	9389.1561	**4629.3282**
		44066	52845.5524	*18803.2937*	**13335.5235**
	R^2 (↑)	44005	0.5541	*0.7494*	**0.7692**
		44008	0.5290	**0.7216**	*0.7205*
		44011	0.3318	**0.4563**	*0.4460*
		44020	0.4984	**0.5867**	*0.5904*
		44025	0.7106	*0.7568*	**0.9140**
		44027	0.2871	**0.3299**	*0.3199*
		44065	*0.6445*	0.6232	**0.7890**
		44066	0.3772	*0.7580*	**0.8384**
	MaxSE (↓)	44005	11.2392	*6.4114*	**6.0429**
		44008	5.9840	**3.8672**	*4.0236*
		44011	21.2563	**15.5661**	*16.0240*
		44020	**3.9678**	*4.4735*	4.5082
		44025	*3.9641*	6.0605	**1.1741**
		44027	21.6919	**16.2316**	*17.5281*
		44065	2.7312	*2.6608*	**1.3428**
		44066	15.4801	*10.3903*	**10.0580**
	MRE (↓)	44005	0.2940	0.2230	**0.2167**
		44008	0.3212	*0.2225*	**0.2217**
		44011	0.3293	**0.3110**	*0.3167*
		44020	0.3293	**0.2944**	*0.2949*
		44025	0.2388	*0.1957*	**0.1225**
		44027	**0.4052**	*0.4273*	0.4293
		44065	*0.2218*	0.2465	**0.1823**
		44066	0.3397	*0.2051*	**0.1725**

The best results are highlighted in **bold**, whereas the second best are in *italics*.

Table 8
Results achieved on the second phase of the approach, optimising models by MSE. The results of the stochastic models are expressed as their mean and Standard Deviation (*SD*): $Mean_{SD}$.

	Metric	Buoy	LinearRegression	Ridge	Lasso	ElasticNet	SVR	MLP	RandomForest	PU	SU	PU+SU
Gulf of Alaska	MSE (↓)	46061	0.1581	0.1566	0.1580	0.1566	0.1811	$0.1560_{0.0026}$	$0.1612_{0.0025}$	*$0.1528_{0.0015}$*	$0.1535_{0.0032}$	**$0.1499_{0.0014}$**
		46076	0.1745	0.1744	0.1743	0.1743	0.1736	$0.1703_{0.0019}$	$0.1756_{0.0029}$	*$0.1697_{0.0006}$*	**$0.1688_{0.0016}$**	**$0.1688_{0.0022}$**
		46078	**0.4452**	**0.4452**	*0.4455*	*0.4455*	0.4532	$0.4457_{0.0006}$	$0.4524_{0.0008}$	$0.4460_{0.0005}$	$0.4478_{0.0025}$	$0.4466_{0.0007}$
		46082	0.3655	0.3655	0.3651	0.3651	0.3766	$0.3587_{0.0057}$	$0.3556_{0.0011}$	$0.3547_{0.0028}$	*$0.3541_{0.0023}$*	**$0.3480_{0.0025}$**
		46085	0.4428	0.4428	0.4424	0.4424	0.4382	$0.4370_{0.0015}$	$0.4486_{0.0018}$	*$0.4351_{0.0013}$*	**$0.4342_{0.0019}$**	$0.4377_{0.0032}$
	SERA (↓)	46061	10664.4483	11148.2822	10759.1228	10849.2567	14727.6876	$10767.3642_{328.3058}$	$11461.6720_{246.4244}$	*$6564.0699_{73.6302}$*	$6807.1688_{175.8786}$	**$6538.4622_{100.7926}$**
		46076	15717.2647	15716.4361	15803.4999	15802.4383	15367.2084	$15115.2904_{273.2186}$	$15804.6627_{335.0880}$	$8518.9647_{44.5683}$	*$8435.2063_{128.3628}$*	**$8398.3464_{185.7868}$**
		46078	29039.6746	29039.8674	29170.9211	29170.9211	29754.7437	$28867.6146_{387.1405}$	$29010.5159_{76.1822}$	**$18394.7586_{104.3305}$**	$18470.6403_{233.2760}$	*$18455.2227_{89.3421}$*
		46082	17881.0046	17881.0725	17940.9341	17940.9341	18892.7873	$17130.3207_{463.0613}$	$16835.6884_{93.4771}$	$10278.7888_{194.9226}$	*$9867.1941_{206.0941}$*	**$9739.1154_{152.4571}$**
		46085	26258.4783	26258.6329	26383.4965	26383.4965	27051.5767	$25662.5181_{288.7405}$	$25422.8782_{166.1083}$	*$15217.4397_{95.4788}$*	**$14935.6931_{114.5193}$**	$15231.3725_{298.5966}$
	R^2 (↑)	46061	0.8498	0.8514	0.8496	0.8508	0.8282	$0.8517_{0.0025}$	$0.8474_{0.0030}$	*$0.8556_{0.0013}$*	$0.8548_{0.0027}$	**$0.8588_{0.0013}$**
		46076	0.8709	0.8710	0.8709	0.8709	0.8717	$0.8745_{0.0008}$	$0.8697_{0.0021}$	$0.8746_{0.0003}$	*$0.8756_{0.0011}$*	**$0.8757_{0.0016}$**
		46078	**0.7194**	**0.7194**	**0.7194**	**0.7194**	0.7167	**$0.7194_{0.0002}$**	$0.7147_{0.0005}$	*$0.7188_{0.0003}$*	$0.7177_{0.0016}$	$0.7184_{0.0004}$
		46082	0.7691	0.7691	0.7691	0.7691	0.7668	$0.7725_{0.0037}$	$0.7730_{0.0007}$	$0.7759_{0.0015}$	*$0.7760_{0.0015}$*	**$0.7790_{0.0013}$**
		46085	0.7910	0.7910	0.7911	0.7911	*0.7947*	$0.7936_{0.0007}$	$0.7881_{0.0008}$	$0.7945_{0.0006}$	**$0.7950_{0.0008}$**	$0.7934_{0.0015}$
	MaxSE (↓)	46061	5.7438	6.2272	5.8374	5.9886	8.8995	$5.8612_{0.1636}$	**$4.8759_{0.1885}$**	$5.5329_{0.0753}$	*$5.1899_{0.1370}$*	$5.3058_{0.1358}$
		46076	11.3678	11.3822	11.4993	11.5069	*10.6332*	$11.3026_{0.1766}$	**$10.2167_{0.4996}$**	$11.1380_{0.1235}$	$11.0649_{0.3006}$	$10.9349_{0.3733}$
		46078	**15.6620**	*15.6622*	15.8154	15.8154	16.9998	$15.8207_{0.2826}$	$16.3259_{0.2980}$	$16.4247_{0.1945}$	$16.6703_{0.6856}$	$16.5127_{0.1759}$
		46082	7.1700	7.1699	7.1677	7.1677	7.5273	*$7.1135_{0.1143}$*	$7.3182_{0.1269}$	$7.2356_{0.1366}$	$7.1569_{0.1879}$	**$7.0188_{0.1963}$**
		46085	8.7660	8.7660	8.8144	8.8144	8.9335	**$8.5597_{0.1109}$**	$8.7415_{0.1381}$	*$8.6043_{0.0433}$*	$8.7744_{0.2632}$	$8.6100_{0.1298}$
	MRE (↓)	46061	0.1665	0.1655	0.1663	0.1655	0.1676	$0.1651_{0.0016}$	$0.1667_{0.0013}$	$0.1631_{0.0008}$	*$0.1621_{0.0017}$*	**$0.1599_{0.0007}$**
		46076	0.1506	0.1506	0.1502	0.1502	0.1504	$0.1498_{0.0009}$	$0.1527_{0.0012}$	*$0.1493_{0.0005}$*	**$0.1491_{0.0007}$**	$0.1493_{0.0008}$
		46078	0.2034	0.2034	0.2038	0.2038	**0.2014**	$0.2035_{0.0009}$	$0.2056_{0.0003}$	*$0.2024_{0.0003}$*	$0.2034_{0.0012}$	$0.2029_{0.0003}$
		46082	0.1756	0.1756	0.1755	0.1755	0.1783	$0.1746_{0.0012}$	$0.1755_{0.0003}$	*$0.1739_{0.0006}$*	*$0.1739_{0.0006}$*	**$0.1728_{0.0007}$**
		46085	0.1651	0.1651	0.1651	0.1651	**0.1629**	$0.1644_{0.0004}$	$0.1683_{0.0004}$	*$0.1637_{0.0003}$*	$0.1641_{0.0005}$	$0.1640_{0.0005}$
	Metric	Buoy	LinearRegression	Ridge	Lasso	ElasticNet	SVR	MLP	RandomForest	PU	SU	PU+SU
Northeast Coast	MSE (↓)	44005	0.1044	0.1045	0.1037	0.1039	0.1044	*$0.1033_{0.0019}$*	$0.1154_{0.0009}$	$0.1053_{0.0044}$	**$0.1031_{0.0034}$**	$0.1117_{0.0110}$
		44008	0.1623	0.1585	0.1566	0.1562	0.1537	*$0.1533_{0.0026}$*	$0.1563_{0.0012}$	**$0.1514_{0.0041}$**	$0.1563_{0.0086}$	$0.1541_{0.0058}$
		44011	0.8833	0.8833	0.8802	0.8802	0.8904	*$0.8341_{0.0096}$*	**$0.8220_{0.0036}$**	$0.8465_{0.0056}$	$0.8433_{0.0232}$	$0.8380_{0.0292}$
		44020	0.0418	0.0415	0.0425	0.0416	0.0416	*$0.0414_{0.0006}$*	$0.0465_{0.0002}$	$0.0804_{0.0907}$	**$0.0412_{0.0007}$**	$0.0419_{0.0008}$
		44025	0.0339	0.0339	0.0342	0.0342	**0.0330**	$0.0334_{0.0005}$	$0.0368_{0.0002}$	$0.0709_{0.1759}$	*$0.0332_{0.0005}$*	**$0.0330_{0.0005}$**
		44027	0.4629	0.4628	0.4612	0.4610	**0.4520**	$0.4563_{0.0023}$	$0.4539_{0.0009}$	$0.4560_{0.0007}$	$0.4557_{0.0012}$	*$0.4536_{0.0021}$*
		44065	0.0338	0.0338	0.0368	0.0368	0.0347	$0.0353_{0.0005}$	$0.0337_{0.0002}$	*$0.0333_{0.0002}$*	**$0.0330_{0.0008}$**	**$0.0330_{0.0007}$**
		44066	**0.1136**	**0.1136**	0.1193	0.1193	*0.1152*	*$0.1152_{0.0012}$*	$0.1314_{0.0010}$	$0.1158_{0.0024}$	$0.1179_{0.0030}$	$0.1186_{0.0032}$
	SERA (↓)	44005	5605.9115	5612.7849	5633.8913	5650.9320	5975.1325	$5560.0195_{136.8825}$	$6342.4117_{81.3010}$	**$3613.8875_{271.4587}$**	*$3744.3062_{229.8895}$*	$3907.2206_{574.5327}$
		44008	7742.7134	7626.2445	7592.6547	7558.5572	8323.8136	$7503.0712_{196.5287}$	$8369.1406_{89.0931}$	*$4846.1251_{137.6652}$*	$5271.7589_{206.6653}$	**$4835.3828_{231.8689}$**
		44011	28065.6455	28065.1765	28144.1929	28144.1929	29571.0343	$25436.1695_{540.9184}$	$24971.7589_{129.6492}$	**$16669.7033_{473.6500}$**	*$16868.4807_{1423.8097}$*	$17009.1688_{1311.5456}$
		44020	3208.3297	3234.5864	3326.8788	3228.6959	2918.8629	$3127.1180_{91.6059}$	$3230.2454_{23.7052}$	**$1259.9140_{132.9363}$**	*$1796.7355_{42.3087}$*	$1840.4305_{69.4784}$
		44025	2075.4542	2075.4723	2095.8021	2095.8021	1992.2001	$2010.2364_{62.1105}$	$2243.5931_{23.5579}$	$1261.1829_{257.3111}$	*$1172.6459_{70.9526}$*	**$1116.3864_{46.8647}$**
		44027	42511.8428	42479.6281	42085.3125	42048.5222	40747.8902	$40089.7654_{886.2117}$	$38872.1555_{170.8272}$	$25648.5797_{106.1290}$	*$25569.0626_{187.8732}$*	**$25436.7969_{239.5733}$**
		44065	2388.9036	2389.0303	2643.2509	2643.2509	2520.1190	$2522.0780_{49.2890}$	$2382.5633_{19.3104}$	$1537.4463_{18.9930}$	*$1519.2521_{59.6646}$*	**$1506.1100_{50.6394}$**
		44066	8626.3934	8626.7148	9398.2295	9398.2295	9126.3871	$8947.9964_{361.1221}$	$9907.7310_{113.4463}$	**$6118.0188_{301.4930}$**	$6296.0088_{273.4421}$	*$6252.8932_{235.3900}$*
	R^2 (↑)	44005	0.8825	0.8824	0.8828	0.8826	**0.8886**	$0.8848_{0.0018}$	$0.8703_{0.0010}$	$0.8826_{0.0041}$	*$0.8854_{0.0040}$*	$0.8767_{0.0106}$
		44008	0.7572	0.7628	0.7656	0.7661	0.7724	*$0.7750_{0.0025}$*	$0.7722_{0.0016}$	**$0.7754_{0.0055}$**	$0.7716_{0.0125}$	$0.7727_{0.0077}$
		44011	0.5028	0.5028	0.5024	0.5024	0.4967	*$0.5224_{0.0032}$*	**$0.5265_{0.0014}$**	$0.5142_{0.0038}$	$0.5183_{0.0134}$	$0.5159_{0.0142}$
		44020	0.6036	0.6066	0.5969	0.6056	*0.6165*	$0.6086_{0.0051}$	$0.5721_{0.0020}$	**$0.8788_{0.1111}$**	$0.6105_{0.0061}$	$0.6048_{0.0074}$
		44025	0.9400	0.9400	0.9398	0.9398	**0.9419**	$0.9411_{0.0005}$	$0.9351_{0.0003}$	$0.9119_{0.1159}$	*$0.9413_{0.0009}$*	**$0.9419_{0.0009}$**
		44027	0.3440	0.3440	0.3442	0.3442	**0.3574**	$0.3538_{0.0012}$	$0.3560_{0.0009}$	$0.3557_{0.0006}$	$0.3554_{0.0009}$	*$0.3569_{0.0011}$*
		44065	0.8863	0.8863	0.8758	0.8758	0.8830	$0.8814_{0.0018}$	$0.8867_{0.0005}$	$0.8881_{0.0007}$	*$0.8890_{0.0028}$*	**$0.8891_{0.0023}$**
		44066	**0.8801**	**0.8801**	0.8771	0.8771	*0.8792*	$0.8790_{0.0012}$	$0.8613_{0.0010}$	$0.8788_{0.0014}$	$0.8764_{0.0030}$	$0.8761_{0.0036}$
	MaxSE (↓)	44005	3.6107	3.5943	3.4342	*3.4161*	3.7062	$3.5888_{0.2252}$	**$3.3224_{0.1877}$**	$4.5062_{1.5749}$	$3.8340_{0.7367}$	$4.5098_{1.2862}$
		44008	3.9266	3.9518	*3.8959*	3.9378	4.1340	**$3.7391_{0.0956}$**	$4.7746_{0.6225}$	$4.2322_{3.3110}$	$3.9289_{0.2078}$	$4.4385_{3.8965}$
		44011	13.3153	*13.3151*	13.3727	13.3727	14.1031	$13.5383_{0.2246}$	$13.9468_{0.3843}$	**$12.8397_{0.1075}$**	$14.5254_{3.4158}$	$13.7603_{3.4804}$
		44020	4.2815	*4.2657*	**4.2636**	4.2723	4.2845	$4.2880_{0.0622}$	$4.4073_{0.0321}$	$67.8355_{132.0093}$	$4.3212_{0.0786}$	$4.3301_{0.0906}$
		44025	0.7949	0.7949	0.6894	0.6894	**0.6081**	*$0.6273_{0.0881}$*	$1.3114_{0.1624}$	$53.7501_{256.3741}$	$0.6409_{0.2221}$	$0.6774_{0.4113}$
		44027	19.9385	19.9290	19.8339	19.8283	19.6555	$19.2461_{0.2190}$	**$18.5265_{0.1299}$**	$19.3549_{0.0728}$	*$19.1377_{0.1745}$*	$19.2003_{0.1412}$
		44065	1.0504	1.0503	**0.8769**	**0.8769**	1.0618	$0.9643_{0.0291}$	*$0.9584_{0.0471}$*	$1.0901_{0.0285}$	$1.1201_{0.1149}$	$1.1451_{0.0900}$
		44066	*7.1190*	7.1209	9.2519	9.2519	9.1340	$7.5839_{0.6516}$	$11.8599_{0.7373}$	**$6.9623_{0.7734}$**	$7.7921_{1.4113}$	$7.6959_{1.3778}$
	MRE (↓)	44005	0.1491	0.1491	0.1484	0.1485	*0.1468*	$0.1478_{0.0013}$	$0.1570_{0.0006}$	$0.1475_{0.0027}$	**$0.1467_{0.0020}$**	$0.1515_{0.0060}$
		44008	0.1942	0.1915	0.1892	0.1897	*0.1835*	$0.1858_{0.0030}$	$0.1840_{0.0006}$	**$0.1814_{0.0013}$**	$0.1852_{0.0070}$	$0.1856_{0.0039}$
		44011	0.2878	0.2878	0.2876	0.2876	0.2889	*$0.2812_{0.0019}$*	**$0.2802_{0.0006}$**	$0.2827_{0.0010}$	$0.2814_{0.0030}$	$0.2813_{0.0032}$
		44020	0.2909	0.2909	0.2943	0.2909	**0.2872**	$0.2902_{0.0028}$	$0.3033_{0.0009}$	$0.2922_{0.0030}$	*$0.2897_{0.0031}$*	$0.2915_{0.0031}$
		44025	0.1027	0.1027	0.1043	0.1043	*0.1013*	$0.1017_{0.0011}$	$0.1068_{0.0002}$	$0.1053_{0.0037}$	**$0.1011_{0.0007}$**	**$0.1011_{0.0007}$**
		44027	0.3888	0.3889	0.3896	0.3896	**0.3843**	$0.3888_{0.0011}$	$0.3897_{0.0004}$	$0.3894_{0.0007}$	*$0.3883_{0.0010}$*	$0.3884_{0.0021}$
		44065	0.1264	0.1264	0.1320	0.1320	0.1275	$0.1292_{0.0011}$	$0.1255_{0.0002}$	$0.1255_{0.0004}$	**$0.1250_{0.0015}$**	*$0.1251_{0.0012}$*
		44066	*0.1471*	*0.1471*	0.1498	0.1498	**0.1465**	$0.1471_{0.0017}$	$0.1574_{0.0005}$	$0.1485_{0.0012}$	$0.1492_{0.0019}$	$0.1496_{0.0015}$

The best result is highlighted in **bold**; the second one best result is shown in *italics*.

Table 9
Results achieved on the second phase of the approach, optimising models by SERA. The results of the stochastic models are expressed as their mean and Standard Deviation (*SD*): $Mean_{SD}$.

	Metric	Buoy	LinearRegression	Ridge	Lasso	ElasticNet	SVR	MLP	RandomForest	PU	SU	PU+SU
Gulf of Alaska	MSE (↓)	46061	0.1581	*0.1576*	0.1580	**0.1566**	0.1811	$0.1579_{0.0029}$	$0.1632_{0.0009}$	$0.1955_{0.0047}$	$0.1886_{0.0063}$	$0.1845_{0.0089}$
		46076	0.1745	0.1745	0.1743	0.1743	*0.1725*	$\mathbf{0.1703_{0.0021}}$	$0.1807_{0.0067}$	$0.1918_{0.0065}$	$0.1954_{0.0070}$	$0.2000_{0.0087}$
		46078	**0.4452**	**0.4452**	*0.4455*	*0.4455*	0.4532	$\mathit{0.4455_{0.0004}}$	$0.4533_{0.0024}$	$0.9506_{0.0102}$	$0.9442_{0.0071}$	$0.9392_{0.0136}$
		46082	0.3655	0.3655	0.3651	0.3651	0.3766	$\mathit{0.3594_{0.0059}}$	$\mathbf{0.3561_{0.0016}}$	$0.4638_{0.0100}$	$0.4783_{0.0192}$	$0.4824_{0.0235}$
		46085	0.4428	0.4428	0.4424	0.4424	*0.4382*	$\mathbf{0.4370_{0.0013}}$	$0.4466_{0.0031}$	$0.6729_{0.0143}$	$0.6901_{0.0135}$	$0.7039_{0.0298}$
	SERA (↓)	46061	10664.4483	10689.0179	10759.1228	10849.2567	14727.6876	$10753.1888_{240.4077}$	$11630.7253_{139.5485}$	$\mathit{6528.1300_{68.8213}}$	$6549.7040_{93.1419}$	$\mathbf{6362.0409_{188.1432}}$
		46076	15717.2647	15717.2553	15803.4999	15803.4999	14817.1561	$15110.9463_{282.0997}$	$16125.3902_{546.3187}$	$7435.9183_{102.6282}$	$\mathit{7293.7648_{69.0909}}$	$\mathbf{7252.7616_{125.3626}}$
		46078	29039.6746	29039.8674	29170.9211	29170.9211	29754.7437	$28737.8139_{350.7998}$	$29058.4763_{198.3715}$	$\mathbf{13675.3231_{17.7498}}$	$13705.4455_{29.6155}$	$\mathit{13687.1571_{24.8219}}$
		46082	17881.0046	17881.0725	17940.9341	17940.9341	18892.7873	$17130.4101_{390.5689}$	$16873.6626_{135.6294}$	$\mathbf{7594.4811_{17.5420}}$	$7790.6611_{113.7059}$	$\mathit{7668.2098_{104.0323}}$
		46085	26258.4783	26258.6329	26383.4965	26383.4965	27051.5767	$25610.0824_{314.6427}$	$25189.5615_{219.8686}$	$\mathbf{11765.2626_{82.1648}}$	$11881.4101_{154.1414}$	$\mathit{11875.6736_{144.0537}}$
	R^2 (↑)	46061	0.8498	0.8501	0.8496	*0.8508*	0.8282	$0.8498_{0.0028}$	$0.8448_{0.0008}$	$0.8452_{0.0018}$	$0.8499_{0.0027}$	$\mathbf{0.8515_{0.0054}}$
		46076	0.8709	0.8709	0.8709	0.8709	*0.8743*	$\mathbf{0.8747_{0.0006}}$	$0.8662_{0.0047}$	$0.8696_{0.0031}$	$0.8718_{0.0031}$	$0.8727_{0.0057}$
		46078	**0.7194**	**0.7194**	**0.7194**	**0.7194**	0.7167	$\mathit{0.7193_{0.0002}}$	$0.7141_{0.0015}$	$0.6879_{0.0042}$	$0.6820_{0.0018}$	$0.6866_{0.0082}$
		46082	0.7691	0.7691	0.7691	0.7691	0.7668	$\mathit{0.7722_{0.0037}}$	$\mathbf{0.7727_{0.0010}}$	$0.7704_{0.0036}$	$0.7639_{0.0082}$	$0.7665_{0.0071}$
		46085	0.7910	0.7910	0.7911	0.7911	**0.7947**	$\mathit{0.7936_{0.0006}}$	$0.7891_{0.0014}$	$0.7883_{0.0032}$	$0.7827_{0.0042}$	$0.7854_{0.0072}$
	MaxSE (↓)	46061	5.7438	5.7927	5.8374	5.9886	8.8995	$5.7866_{0.1065}$	$5.0186_{0.1632}$	$5.0635_{0.4525}$	$\mathbf{4.4701_{0.2551}}$	$\mathit{4.5472_{0.3304}}$
		46076	11.3678	11.3679	11.4993	11.4993	9.2493	$11.2829_{0.1603}$	$10.2451_{0.5267}$	$9.4982_{0.1847}$	$\mathit{8.9685_{0.1928}}$	$\mathbf{8.8731_{0.3735}}$
		46078	15.6620	15.6622	15.8154	15.8154	16.9998	$15.7940_{0.2193}$	$16.2940_{0.4009}$	$\mathbf{12.7660_{0.0811}}$	$12.8762_{0.2287}$	$\mathit{12.8489_{0.1391}}$
		46082	7.1700	7.1699	7.1677	7.1677	7.5273	$7.1111_{0.0905}$	$7.3453_{0.1566}$	$\mathit{5.6095_{0.1156}}$	$5.7074_{0.4792}$	$\mathbf{5.4034_{0.3683}}$
		46085	8.7660	8.7660	8.8144	8.8144	8.9335	$8.5407_{0.1291}$	$8.9337_{0.1988}$	$\mathbf{6.4433_{0.9217}}$	$6.8300_{0.1869}$	$\mathit{6.6228_{0.2882}}$
	MRE (↓)	46061	0.1665	**0.1662**	0.1663	0.1655	0.1676	$\mathit{0.1662_{0.0018}}$	$0.1678_{0.0004}$	$0.1950_{0.0062}$	$0.1898_{0.0076}$	$0.1888_{0.0078}$
		46076	0.1506	0.1506	*0.1502*	*0.1502*	0.1507	$\mathbf{0.1497_{0.0011}}$	$0.1551_{0.0031}$	$0.1673_{0.0048}$	$0.1741_{0.0076}$	$0.1803_{0.0093}$
		46078	0.2034	0.2034	0.2038	0.2038	**0.2014**	$\mathit{0.2027_{0.0009}}$	$0.2058_{0.0006}$	$0.3495_{0.0030}$	$0.3434_{0.0018}$	$0.3440_{0.0044}$
		46082	0.1756	0.1756	*0.1755*	*0.1755*	0.1783	$\mathbf{0.1747_{0.0013}}$	$0.1757_{0.0005}$	$0.2196_{0.0030}$	$0.2216_{0.0067}$	$0.2258_{0.0099}$
		46085	0.1651	0.1651	0.1651	0.1651	**0.1629**	$\mathit{0.1644_{0.0005}}$	$0.1676_{0.0005}$	$0.2264_{0.0040}$	$0.2279_{0.0041}$	$0.2332_{0.0078}$
	Metric	Buoy	LinearRegression	Ridge	Lasso	ElasticNet	SVR	MLP	RandomForest	PU	SU	PU+SU
Northeast Coast	MSE (↓)	44005	0.1044	0.1049	**0.1037**	0.1050	0.1044	$\mathit{0.1039_{0.0022}}$	$0.1155_{0.0010}$	$0.1188_{0.0086}$	$0.1209_{0.0081}$	$0.1199_{0.0098}$
		44008	0.1623	0.1585	0.1566	0.1566	*0.1537*	$\mathbf{0.1527_{0.0026}}$	$0.1565_{0.0016}$	$0.7886_{2.6638}$	$0.3378_{0.0872}$	$0.3215_{0.0768}$
		44011	0.8833	0.8833	0.8802	0.8802	0.8904	$\mathit{0.8367_{0.0093}}$	$\mathbf{0.8222_{0.0078}}$	$1.4533_{0.0365}$	$1.5241_{0.0886}$	$1.5010_{1.5010}$
		44020	0.0418	0.0417	0.0425	0.0416	**0.0398**	$\mathit{0.0415_{0.0007}}$	$0.0466_{0.0004}$	$0.0758_{0.0082}$	$0.0729_{0.0038}$	$0.0744_{0.0047}$
		44025	0.0339	0.0339	0.0342	0.0342	**0.0330**	$\mathit{0.0336_{0.0005}}$	$0.0368_{0.0002}$	$0.0468_{0.0121}$	$0.0438_{0.0027}$	$0.0438_{0.0025}$
		44027	0.4629	0.4629	0.4612	0.4612	**0.4557**	$\mathit{0.4559_{0.0018}}$	$0.4564_{0.0016}$	$0.5685_{0.0077}$	$0.5643_{0.0099}$	$0.5659_{0.0136}$
		44065	**0.0338**	**0.0338**	0.0368	0.0368	*0.0347*	$0.0354_{0.0005}$	$\mathbf{0.0338_{0.0003}}$	$0.0434_{0.0015}$	$0.0433_{0.0019}$	$0.0424_{0.0027}$
		44066	**0.1136**	**0.1136**	0.1193	0.1193	*0.1152*	$0.1153_{0.0014}$	$0.1313_{0.0019}$	$0.1648_{0.0079}$	$0.1859_{0.0191}$	$0.1723_{0.0131}$
	SERA (↓)	44005	5605.9115	5681.2963	5633.8913	5738.1196	5975.1325	$5601.8064_{152.2240}$	$6343.8871_{85.8261}$	$\mathbf{3482.6442_{218.9006}}$	$3560.3486_{244.3136}$	$\mathit{3528.1412_{370.5786}}$
		44008	7742.7134	7626.2445	7592.6547	7592.6547	8323.8136	$7531.9970_{215.9654}$	$8374.2398_{130.8725}$	$\mathit{5267.8654_{2523.6104}}$	$5275.3799_{485.7511}$	$\mathbf{5113.9226_{500.1604}}$
		44011	28065.6455	28065.1765	28144.1929	28144.1929	29571.0343	$25565.6633_{498.9032}$	$25634.7770_{459.4333}$	$\mathbf{11394.5672_{130.4803}}$	$13000.5643_{2028.2205}$	$\mathit{11752.9770_{11752.9770}}$
		44020	3208.3297	3207.3917	3326.8788	3228.6959	2800.7833	$3100.5555_{85.3141}$	$3239.6479_{36.0648}$	$\mathbf{1313.7564_{38.3863}}$	$1326.6046_{27.1161}$	$\mathit{1319.3581_{35.3638}}$
		44025	2075.4542	2075.4723	2095.8021	2095.8021	1992.2001	$2012.8634_{67.9121}$	$2245.7172_{30.9393}$	$\mathit{1124.1819_{29.6539}}$	$1127.3527_{47.4385}$	$\mathbf{1099.3406_{42.0522}}$
		44027	42511.8428	42511.5180	42085.3125	42085.3125	39802.5155	$40033.4009_{770.9741}$	$38834.4376_{178.8641}$	$18035.6725_{77.6232}$	$\mathit{17908.8358_{299.8570}}$	$\mathbf{17739.7083_{350.6746}}$
		44065	2388.9036	2389.0303	2643.2509	2643.2509	2520.1190	$2530.4941_{52.3965}$	$2386.9819_{28.3249}$	$1431.9683_{24.4679}$	$\mathbf{1408.6447_{23.6190}}$	$\mathit{1417.7112_{27.6952}}$
		44066	8626.3934	8626.7148	9398.2295	9398.2295	9126.3871	$8929.5531_{369.3166}$	$9947.5424_{155.4176}$	$\mathit{5616.5070_{220.0157}}$	$5680.5942_{242.2960}$	$\mathbf{5584.9080_{267.1034}}$
	R^2 (↑)	44005	0.8825	0.8818	0.8828	0.8814	**0.8886**	$\mathit{0.8840_{0.0024}}$	$0.8703_{0.0011}$	$0.8732_{0.0110}$	$0.8721_{0.0074}$	$0.8733_{0.0091}$
		44008	0.7572	0.7628	0.7656	0.7656	*0.7724*	$\mathbf{0.7749_{0.0030}}$	$0.7719_{0.0019}$	$0.6966_{0.1342}$	$0.7099_{0.0544}$	$0.7244_{0.0539}$
		44011	0.5028	0.5028	0.5024	0.5024	0.4967	$0.5217_{0.0035}$	$\mathit{0.5240_{0.0041}}$	$\mathbf{0.5264_{0.0075}}$	$0.4709_{0.0314}$	$0.5145_{0.5144}$
		44020	0.6036	0.6043	0.5969	0.6056	**0.6268**	$\mathit{0.6086_{0.0036}}$	$0.5712_{0.0029}$	$0.5847_{0.0444}$	$0.6016_{0.0091}$	$0.5968_{0.0122}$
		44025	0.9400	0.9400	0.9398	0.9398	**0.9419**	$\mathit{0.9409_{0.0006}}$	$0.9350_{0.0004}$	$0.9318_{0.0184}$	$0.9369_{0.0020}$	$0.9387_{0.0035}$
		44027	0.3440	0.3440	0.3442	0.3442	0.3552	$0.3546_{0.0006}$	$0.3540_{0.0014}$	$0.3565_{0.0025}$	$\mathit{0.3589_{0.0029}}$	$\mathbf{0.3598_{0.0027}}$
		44065	0.8863	0.8863	0.8758	0.8758	0.8830	$0.8810_{0.0018}$	$0.8865_{0.0009}$	$\mathbf{0.8870_{0.0023}}$	$0.8868_{0.0025}$	$\mathit{0.8869_{0.0024}}$
		44066	**0.8801**	**0.8801**	0.8771	0.8771	*0.8792*	$0.8791_{0.0014}$	$0.8616_{0.0025}$	$0.8593_{0.0114}$	$0.8641_{0.0126}$	$0.8659_{0.0089}$
	MaxSE (↓)	44005	3.6107	3.4636	3.4342	*3.3339*	3.7062	$3.6012_{0.2357}$	$\mathbf{3.3305_{0.1777}}$	$4.3951_{1.0163}$	$4.5183_{0.9597}$	$4.8000_{1.6630}$
		44008	3.9266	3.9518	*3.8959*	*3.8959*	4.1340	$\mathbf{3.7987_{0.1053}}$	$4.9187_{0.9795}$	$489.1272_{2564.7203}$	$5.7118_{3.9921}$	$7.1985_{7.8524}$
		44011	*13.3153*	**13.3151**	13.3727	13.3727	14.1031	$13.5408_{0.1968}$	$13.6670_{0.2713}$	$17.1146_{1.1567}$	$17.6500_{7.4322}$	$18.3466_{18.3466}$
		44020	4.2815	4.2818	4.2636	4.2723	4.3005	$4.3477_{0.0571}$	$4.3959_{0.0419}$	$6.9224_{9.3477}$	$\mathit{3.7389_{0.1769}}$	$\mathbf{3.7366_{0.2224}}$
		44025	0.7949	0.7949	0.6894	0.6894	*0.6081*	$0.6304_{0.0920}$	$1.3083_{0.1730}$	$4.6355_{16.0142}$	$\mathbf{0.5840_{0.0962}}$	$0.6379_{0.3755}$
		44027	19.9385	19.9384	19.8339	19.8339	18.5037	$19.2512_{0.2130}$	$18.4604_{0.2434}$	$15.2201_{0.0638}$	$\mathit{14.9912_{0.3099}}$	$\mathbf{14.8523_{0.3427}}$
		44065	1.0504	1.0503	**0.8769**	**0.8769**	1.0618	$\mathit{0.9624_{0.0274}}$	$0.9744_{0.0686}$	$1.3645_{0.0543}$	$1.4717_{0.0557}$	$1.4855_{0.0546}$
		44066	7.1190	7.1209	9.2519	9.2519	9.1340	$7.8944_{0.7870}$	$11.8722_{1.0696}$	$\mathbf{6.3533_{0.7791}}$	$7.0904_{1.2825}$	$\mathit{6.9607_{2.2956}}$
	MRE (↓)	44005	0.1491	0.1493	0.1484	0.1491	**0.1468**	$\mathit{0.1483_{0.0016}}$	$0.1570_{0.0007}$	$0.1679_{0.0055}$	$0.1703_{0.0077}$	$0.1699_{0.0087}$
		44008	0.1942	0.1915	0.1892	0.1892	*0.1835*	$0.1850_{0.0028}$	$\mathbf{0.1812_{0.0009}}$	$0.2883_{0.0289}$	$0.3221_{0.0553}$	$0.3117_{0.0371}$
		44011	0.2878	0.2878	0.2876	0.2876	0.2889	$\mathit{0.2818_{0.0018}}$	$\mathbf{0.2796_{0.0011}}$	$0.4054_{0.0061}$	$0.4153_{0.0114}$	$0.4140_{0.0209}$
		44020	0.2909	0.2907	0.2943	0.2909	**0.2837**	$\mathit{0.2896_{0.0026}}$	$0.3036_{0.0012}$	$0.4306_{0.0144}$	$0.4250_{0.0173}$	$0.4302_{0.0215}$
		44025	0.1027	0.1027	0.1043	0.1043	**0.1013**	$\mathit{0.1021_{0.0012}}$	$0.1069_{0.0003}$	$0.1254_{0.0069}$	$0.1251_{0.0068}$	$0.1261_{0.0064}$
		44027	0.3888	0.3888	0.3896	0.3896	**0.3875**	$\mathit{0.3884_{0.0012}}$	$0.3905_{0.0006}$	$0.4761_{0.0068}$	$0.4721_{0.0114}$	$0.4744_{0.0147}$
		44065	**0.1264**	**0.1264**	0.1320	0.1320	*0.1275*	$0.1294_{0.0010}$	$0.1256_{0.0005}$	$0.1514_{0.0048}$	$0.1506_{0.0072}$	$0.1476_{0.0096}$
		44066	*0.1471*	*0.1471*	0.1498	0.1498	**0.1465**	$0.1473_{0.0020}$	$0.1572_{0.0013}$	$0.1981_{0.0073}$	$0.2191_{0.0149}$	$0.2072_{0.0133}$

The best result is highlighted in **bold**; the second one best result is shown in *italics*.

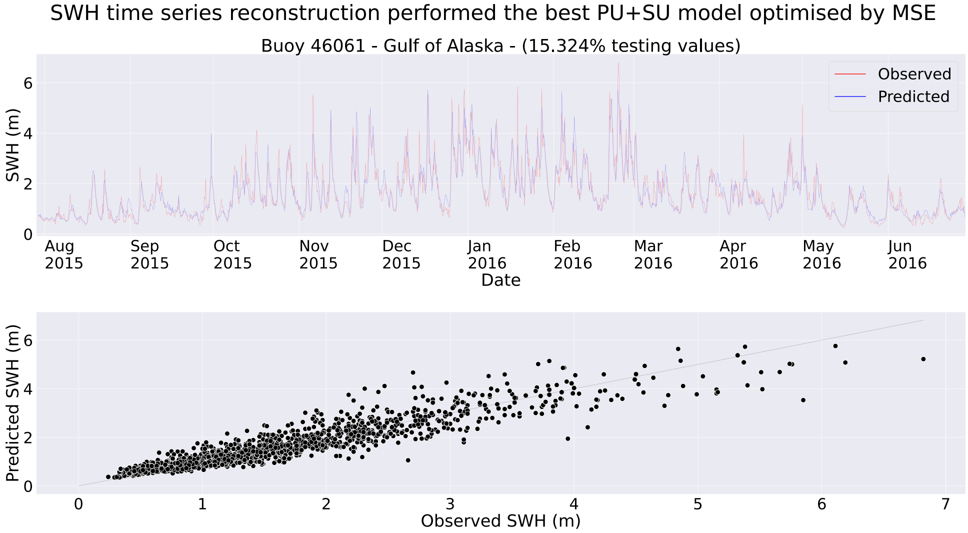

Fig. 7. Comparison of observed (red) vs. predicted (blue) values and scatter plot corresponding to the best PU+SU model optimised by MSE, for the test set of the buoy 46061 of the Gulf of Alaska.

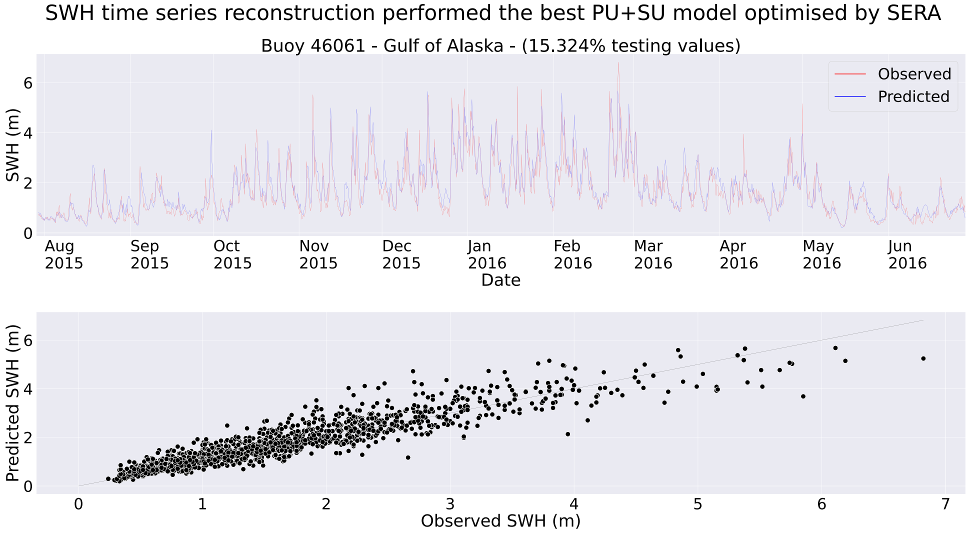

Fig. 8. Comparison of observed (red) vs. predicted (blue) values and scatter plot corresponding to the best PU model optimised by SERA, for the test set of the buoy 46061 of the Gulf of Alaska.

one buoy in the Gulf of Alaska (46078), which, as with the two previous ones, this particular buoy only uses 1 buoy as input. Note that these three buoys have particular weather conditions, being significantly different from the rest of the buoys of the zone. Furthermore, the MRE results reflect that the reconstructions differ from the ground truth by a 10.11% for the buoy achieving the best results (44025) and by a 38.43% for the most challenging buoy to be recovered (44027). Bear in mind that EANNs have been run 30 times (as well as MLP and RandomForest). Since these are average percentages, the best model can reduce this error considerably. Finally, indicate that 9 out of 13 buoys only differ, at maximum, from the ground truth, by a 20%, which demonstrates that the proposed methodology achieves a very good performance when optimised by MSE.

On the other hand, optimisation using SERA is also of great interest, even more than the optimisation using MSE. In this sense, SERA does not only focus on the overall behaviour of the SWH time series as MSE does (i.e. the same attention is paid to both around-mean and extreme values, although the latter are of more interest than the mean values), but pays special attention to these extreme values. Thus, Table 9 shows the results obtained in the second phase when the techniques are optimised by SERA. As can be checked, the best results in terms of SERA are always obtained by EANNs for all the buoys under study. More specifically, PU models obtained the best results for 6 buoys, SU models for 1 buoy, and PU+SU models in the remaining 6 buoys. These excellent results underline the importance of using EANN models with PUs in the hidden layer, either alone (for easier buoys) or combined with SUs (for more complex buoys). It is noteworthy that EANN models not only obtain the best results in terms of SERA but also the second best ones, as happened when MSE was used as the optimisation metric. Therefore, these results demonstrate that the proposed EANNs also achieved an excellent performance when optimised by SERA, contrary to the other techniques applied.

Regarding the remaining performance measures, in the case of MSE and MRE, all the best results are achieved by the state-of-the-art techniques. This behaviour means that these techniques can only recover values around the mean instead of paying special attention to the extreme values. Something similar happens in the

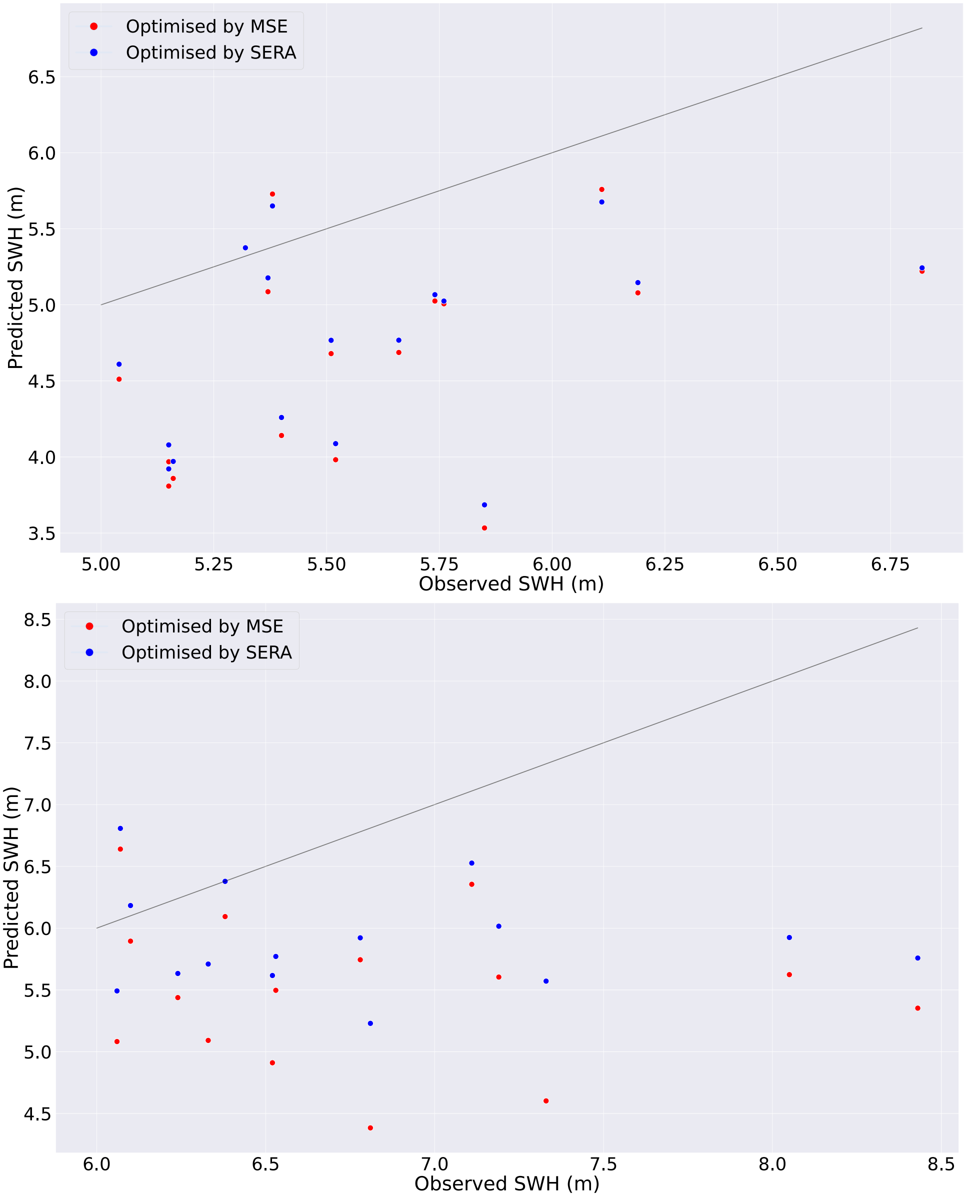

Fig. 9. Scatter plots of the best PU+SU models optimised by MSE (red) vs. by SERA (blue) for the extreme values of the test set of the buoy 46061 (upper) and 46076 (bottom) of the Gulf of Alaska (more than 5 m and 6 m, respectively).

case of R^2 since state-of-the-art techniques achieve most of the best results. In terms of MaxSE, EANNs achieve the best results for half of the buoys at the Northeast Coast and all the buoys at the Gulf of Alaska.

In addition, in order to graphically analyse the performance achieved by EANN models when they are optimised either by MSE or SERA, the buoy 46061 from the Gulf of Alaska has been selected to represent observed vs predicted values. The choice of this buoy is twofold: (1) it has a considerable amount of missing data, almost a quarter of the period of time considered (23.357%, equivalent to 1 year and 5 months approximately), and (2) the range of SWH values is wider than for other buoys, between 0.2 m to 6.9 m, thus, it is easier to show the differences between models optimised using MSE and SERA. In this sense, Fig. 7 shows the SWH time series reconstruction performed by the best PU+SU model optimised by MSE. As can be seen, the predictions are generally good, with no major deviations when SWH values are close to the mean. Nevertheless, in the case of values higher than 4m, the predictions are not that accurate due to those SWH values could be considered extreme values. Also, focusing on the scatter plot, it can be observed that predicted values between 0m and 2.5 m are concentrated around the regression line. However, predicted values higher than 2.5 m are more dispersed. The main reason for this behaviour is that MSE focuses on values around the average, hence, ignoring extreme values. On the other hand, Fig. 8 shows the SWH time series reconstruction performed by the best PU+SU model optimised by SERA. Contrary to the previous case, where MSE optimised the model, SWH values lower than 2.5 m are not accurately predicted. Nevertheless, values over 2.5 m are more accurately predicted. This behaviour is also represented in the scatter plot, where values over 2.5 m are closer to the regression line than in Fig. 7.

Moreover, to shed light on the differences in terms of extreme values between the models optimised by MSE and SERA, Fig. 9 shows the scatter plot of the predictions as well as the regression

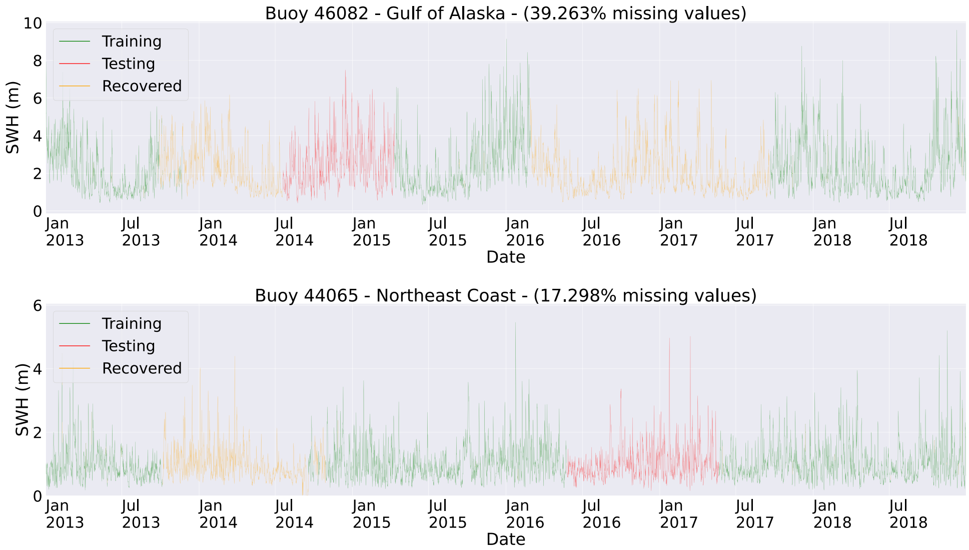

Fig. 10. Recovered SWH time series for the buoys 46082 and 44065 of the Gulf of Alaska and the Northeast Coast, respectively, using MSE as optimisation metric.

lines for two buoys: 46061 (upper) and 46076 (bottom). Specifically, 17 and 15 extreme values of the testing set have been included for buoys 46061 and 46076, i.e. those SWH higher than 5m and 6m, respectively. These amounts represent the 1.25% and the 0.99% of the values belonging to the testing set. Concerning the buoy 46061, the best predictions but one are obtained by the model optimised by SERA, reducing the error performed by a 7.959% on average, i.e. the predictions performed by models optimised by SERA are a 7.959% on average closer to the regression line. In the case of the buoy 46076, the same behaviour is achieved. The best performance over extreme values is obtained by the model optimised by SERA (for 14 out of the 15 values), reducing the error by a 31.873% on average. Summarising, some of the extreme values cannot be estimated as desired. Nevertheless, when SERA optimises the models, a better estimation is achieved for the extreme values.

Furthermore, in order to illustrate the SWH time series after the final recovery, Figs. 10 and 11 show two examples of SWH time series each, where the reconstruction has been performed by the best PU+SU model optimised by MSE and SERA, respectively. Concretely, Fig. 10 shows the reconstructed SWH time series of the buoys 46082 and 44065 of the Gulf of Alaska and the Northeast Coast, respectively. As can be checked, the SWH time series keeps its general behaviour after recovering the missing values, even when they represent a huge portion of the time series, as is the case of the buoy 46082, with more than a 39% of missing values. Note that for the buoy 46082, 19 months are lost in a row (from Mar-2016 to Sep-2017). Nevertheless, the proposed approach is able to recover this gap with high performance. On the other hand, Fig. 11 shows the final reconstructions of the buoys 46061 and 44025 of the Gulf of Alaska and the Northeast Coast, respectively. In this case, it is worthy of mention that in addition to recovering large data gaps, this approach is also good at retrieving intermittent gaps, i.e. missing data is interspersed with training data, as is the case of the buoy 46061 during the year 2013. In general, it can be said that PU+SU models are an excellent approach for recovering missing data, either using MSE or SERA as optimisation metrics, given that they are able to catch the behaviour of the SWH time series (maintaining seasonality and trend among other features of the time series).

Finally, in order to give an insight into the complexity of the models obtained, an analysis in terms of the number of connections is carried out. In this sense, Table 10 shows the mean and standard deviation ($Mean_{SD}$) of the number of connections of the different models applied in the second phase and for each of the buoys of both coastal zones. As can be seen, linear techniques are the ones that require the lowest number of connections, given their simplicity. Among the four linear techniques, LinearReg, Ridge, Lasso and ElasticNet, the last two stand out, given that, for all the buoys either optimised by MSE or by SERA, they have the lowest number of connections, except for the buoy 44008 of the Northeast Coast when optimised by MSE. On the other hand, leaving aside the linear techniques, it can be checked that EANNs using PU, SU or PU+SU basis functions have the lowest number of connections compared to the remaining non-linear techniques, standing out the EANNs with PUs in the hidden layer for being the simplest models.

5.3. Comparison against time series imputation techniques

In order to demonstrate that the proposed approach is able to achieve excellent results, an extra comparison has been included against two state-of-the-art approaches in the time series imputation field: (1) Bidirectional Recurrent Imputation for Time Series (BRITS, [71]) and (2) Self-Attention-based Imputation for Time Series (SAITS, [72]). The reason behind their inclusion is that they are the state-of-the-art in their respective category, i.e. BRITS is the best approach based on Recurrent Neural Networks (RNN), whereas SAITS is the technique based on Self-Attention blocks achieving the best results:

- Bidirectional Recurrent Imputation for Time Series (BRITS, [71]) consists in an RNN approach for missing value imputation in time series data. BRITS imputes the missing values following a bidirectional recurrent dynamical system. The main advantages of this algorithm are: (1) it can manage multiple correlated missing values in time series, (2) it can generalise to time series with nonlinear dynamics underlying, and (3) it provides a data-driven imputation procedure.
  BRITS has been compared against other RNN-based approaches, such as GRU-D [73] and M-RNN [74], outperforming both of them significantly.
- Self-Attention-based Imputation for Time Series (SAITS, [72]) is the most recent approach not only based on Self-Attention mechanism but also tackling the imputation task for time series. Specifically, SAITS is composed of two

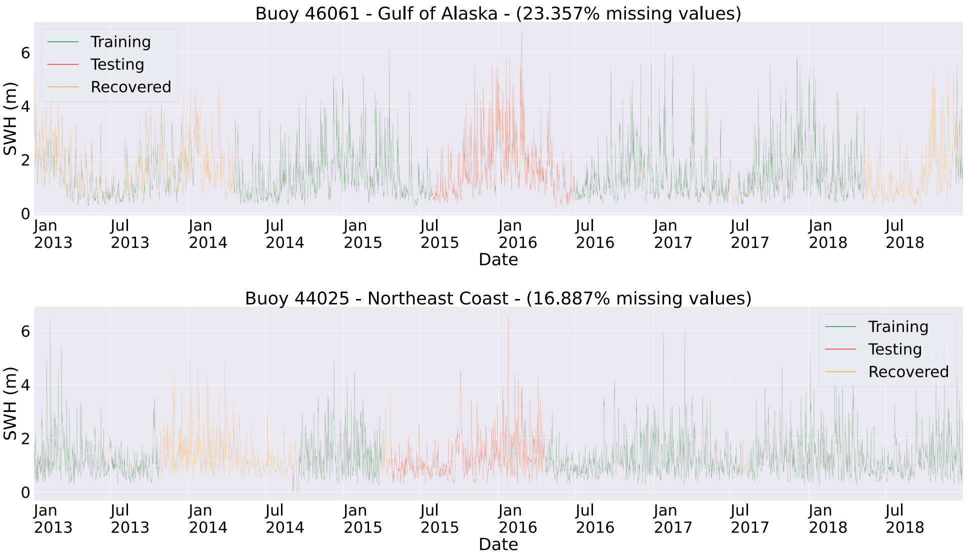

Fig. 11. Recovered SWH time series for the buoys 46061 and 44025 of the Gulf of Alaska and the Northeast Coast, respectively, using SERA as optimisation metric.

Table 10
Comparison between the models of the second phase of the approach optimised by MSE and SERA, respectively, in terms of the number of connections and for each of the buoys of both coastal zones. The number of connections of the stochastic models is expressed as their mean and Standard Deviation (*SD*): $Mean_{SD}$.

		Buoy	LinearReg	Ridge	Lasso	ElasticNet	SVR	MLP	RandomForest	PU	SU	PU+SU
Optimised by MSE	Gulf of Alaska	46061	**3.00**	**3.00**	**3.00**	**3.00**	9584.00	$171.67_{117.92}$	$3727.27_{2164.70}$	$\mathit{6.97}_{\mathit{1.19}}$	$11.90_{0.84}$	$11.23_{0.63}$
		46076	**4.00**	**4.00**	**4.00**	**4.00**	16302.00	$294.03_{107.12}$	$10440.27_{6689.49}$	$\mathit{8.30}_{\mathit{1.70}}$	$12.43_{1.19}$	$12.67_{1.12}$
		46078	**2.00**	**2.00**	**2.00**	**2.00**	8618.00	$117.93_{78.54}$	$3234.33_{1588.14}$	$\mathit{3.33}_{\mathit{0.76}}$	$9.30_{1.51}$	$8.87_{0.57}$
		46082	**4.00**	**4.00**	**4.00**	**4.00**	13390.00	$191.40_{128.11}$	$14519.47_{4596.94}$	$\mathit{5.40}_{\mathit{1.16}}$	$12.23_{1.19}$	$12.13_{0.82}$
		46085	**3.00**	**3.00**	**3.00**	**3.00**	12734.00	$161.63_{99.76}$	$8753.00_{3294.16}$	$\mathit{6.80}_{\mathit{0.89}}$	$11.57_{1.14}$	$10.83_{1.37}$
		Buoy	LinearReg	Ridge	Lasso	ElasticNet	SVR	MLP	RandomForest	PU	SU	PU+SU
	Northeast Coast	44005	7.00	7.00	**6.00**	**6.00**	19056.00	$507.33_{152.76}$	$552305.87_{154594.84}$	$\mathit{6.77}_{\mathit{1.89}}$	$27.33_{2.66}$	$10.93_{2.33}$
		44008	8.00	8.00	*7.00*	*7.00*	21338.00	$471.00_{228.66}$	$14346.27_{4712.38}$	$\mathbf{5.57}_{\mathbf{1.43}}$	$15.03_{1.90}$	$13.87_{2.50}$
		44011	**4.00**	**4.00**	**4.00**	**4.00**	9966.00	$283.93_{112.25}$	$3823.07_{1613.78}$	$\mathit{8.63}_{\mathit{1.22}}$	$19.60_{1.61}$	$10.57_{1.57}$
		44020	**6.00**	**6.00**	**6.00**	**6.00**	20714.00	$430.67_{198.63}$	$787511.80_{231822.80}$	$\mathit{10.67}_{\mathit{1.58}}$	$16.97_{1.83}$	$14.83_{1.97}$
		44025	**5.00**	**5.00**	**5.00**	**5.00**	13037.00	$392.37_{125.99}$	$16776.40_{3626.31}$	$\mathit{7.30}_{\mathit{1.70}}$	$24.37_{1.87}$	$16.57_{1.59}$
		44027	*4.00*	*4.00*	**3.00**	**3.00**	18118.00	$217.70_{112.93}$	$2982.13_{1643.72}$	$7.37_{1.71}$	$10.97_{0.93}$	$16.60_{2.76}$
		44065	*3.00*	*3.00*	**2.00**	**2.00**	8150.00	$268.97_{64.61}$	$13205.00_{5112.76}$	$6.33_{1.32}$	$11.47_{0.78}$	$10.60_{0.62}$
		44066	*7.00*	*7.00*	**3.00**	**3.00**	25797.00	$506.00_{153.24}$	$579712.13_{130831.45}$	$8.63_{1.90}$	$13.90_{1.88}$	$11.77_{2.93}$
		Buoy	LinearReg	Ridge	Lasso	ElasticNet	SVR	MLP	RandomForest	PU	SU	PU+SU
Optimised by SERA	Gulf of Alaska	46061	**3.00**	**3.00**	**3.00**	**3.00**	9584.00	$144.73_{107.39}$	$2840.73_{1634.31}$	$\mathit{10.23}_{\mathit{0.73}}$	$14.87_{0.63}$	$\mathit{10.23}_{\mathit{0.90}}$
		46076	**4.00**	**4.00**	**4.00**	**4.00**	15826.00	$302.47_{102.20}$	$3455.07_{3659.82}$	$12.30_{0.92}$	$16.07_{1.36}$	$\mathit{11.17}_{\mathit{1.37}}$
		46078	**2.00**	**2.00**	**2.00**	**2.00**	8618.00	$150.30_{65.53}$	$2605.13_{1844.75}$	$9.07_{0.37}$	$13.00_{0.00}$	$\mathit{8.93}_{\mathit{0.37}}$
		46082	**4.00**	**4.00**	**4.00**	**4.00**	13390.00	$160.93_{131.68}$	$11782.87_{5760.34}$	$\mathit{11.50}_{\mathit{1.04}}$	$16.67_{1.03}$	$11.73_{1.11}$
		46085	**3.00**	**3.00**	**3.00**	**3.00**	12734.00	$213.00_{100.75}$	$12930.53_{6304.35}$	$10.67_{0.61}$	$15.13_{0.68}$	$\mathit{9.70}_{\mathit{1.02}}$
		Buoy	LinearReg	Ridge	Lasso	ElasticNet	SVR	MLP	RandomForest	PU	SU	PU+SU
	Northeast Coast	44005	*7.00*	*7.00*	**6.00**	**6.00**	19056.00	$424.37_{173.87}$	$544945.80_{174651.31}$	$15.90_{2.16}$	$20.30_{2.09}$	$19.43_{1.98}$
		44008	*8.00*	*8.00*	**7.00**	**7.00**	21338.00	$305.07_{210.07}$	$11580.40_{5634.77}$	$16.07_{2.13}$	$21.80_{2.12}$	$19.90_{2.70}$
		44011	**4.00**	**4.00**	**4.00**	**4.00**	9966.00	$234.67_{110.18}$	$7974.60_{3031.03}$	$\mathit{11.40}_{\mathit{1.04}}$	$16.73_{1.08}$	$14.13_{14.13}$
		44020	**6.00**	**6.00**	**6.00**	**6.00**	21506.00	$389.93_{191.63}$	$767708.27_{277009.53}$	$\mathit{13.83}_{\mathit{1.46}}$	$19.83_{1.66}$	$27.40_{1.75}$
		44025	**5.00**	**5.00**	**5.00**	**5.00**	13037.00	$358.30_{141.38}$	$13471.40_{4825.87}$	$\mathit{12.27}_{\mathit{1.20}}$	$17.77_{1.38}$	$16.03_{1.83}$
		44027	*4.00*	*4.00*	**3.00**	**3.00**	18022.00	$255.00_{115.80}$	$8512.80_{5726.23}$	$16.33_{1.15}$	$23.83_{1.15}$	$15.13_{1.43}$
		44065	*3.00*	*3.00*	**2.00**	**2.00**	8150.00	$261.43_{74.78}$	$10708.87_{6271.59}$	$10.43_{0.68}$	$14.83_{0.70}$	$13.63_{0.61}$
		44066	*7.00*	*7.00*	**3.00**	**3.00**	25797.00	$449.43_{166.98}$	$482345.47_{302117.27}$	$15.83_{1.60}$	$20.90_{1.77}$	$19.07_{2.23}$

The lowest number of connections is highlighted in **bold**; the second one lowest is shown in *italics*.

diagonally-masked self-attention blocks and a weighted combination of both learned representations.
SAITS has been compared against a range of approaches, including BRITS, M-RNN and GP-VAE [75]. SAITS outperforms all these techniques on three well-known datasets of the community.

In addition to these two approaches to sanity check, a comparison against a well-known naïve approach, which consists in recovering all values with the training mean value, has been carried out.

These approaches have been compared against the EANNs, i.e. after carrying out the intermediate reconstruction. Note that these three approaches have been run exactly under the same conditions as EANNs and the rest of ML approaches, namely SVR, LinearRegression, MLP, and so on, i.e. the same inputs and the same number of training patterns, which, in turn, depend on the considered buoy. The BRITS and SAITS approaches have also been run 30 times in order to provide a fair comparison against our approach. Concerning the experimental settings, both approaches have been run using default values.

Therefore, Table 11 shows the results achieved by our approach when optimised by MSE and SERA, respectively, as well as the results achieved by the naïve mean approach and the two state-of-the-art approaches. Note that the results for the EANNs are the same as those of Tables 8 and 9, repeated to enhance the readability. The rest of the ML approaches have not been included, as it was demonstrated that EANNs outperformed them.

As can be observed in Table 11, EANNs optimised by MSE are the best in terms of MSE (12 out of 13 buoys). On the other hand, EANNs evolved by SERA are the best in terms of SERA (11 out of 13). Hence, the results achieved by BRITS and SAITS are not better than those obtained by our approach.

Table 11
Results achieved by our approach when optimised by MSE and SERA in comparison against the naïve mean approach and the two state-of-the-art approaches BRITS and SAITS. The results of the stochastic models are expressed as their mean and Standard Deviation (*SD*): *Mean$_{SD}$*.

	Metric	Buoy	Optimised by MSE			Optimised by SERA			Mean	BRITS	SAITS
			PU	SU	PU+SU	PU	SU	PU+SU			
Gulf of Alaska	MSE (↓)	46061	*0.1528$_{0.0015}$*	0.1535$_{0.0032}$	**0.1499$_{0.0014}$**	0.1955$_{0.0047}$	0.1886$_{0.0063}$	0.1845$_{0.0089}$	1.1286	0.2595$_{0.0034}$	0.4123$_{0.0937}$
		46076	*0.1697$_{0.0006}$*	**0.1688$_{0.0016}$**	**0.1688$_{0.0022}$**	0.1918$_{0.0065}$	0.1954$_{0.0070}$	0.2000$_{0.0087}$	1.3739	0.3597$_{0.0093}$	1.6765$_{0.6726}$
		46078	**0.4460$_{0.0005}$**	0.4478$_{0.0025}$	*0.4466$_{0.0007}$*	0.9506$_{0.0102}$	0.9442$_{0.0071}$	0.9392$_{0.0136}$	1.6032	1.1489$_{0.6015}$	4.0631$_{0.9709}$
		46082	0.3547$_{0.0028}$	*0.3541$_{0.0023}$*	**0.3480$_{0.0025}$**	0.4638$_{0.0100}$	0.4783$_{0.0192}$	0.4824$_{0.0235}$	1.5914	0.5525$_{0.1961}$	0.9547$_{0.2951}$
		46085	*0.4351$_{0.0013}$*	**0.4342$_{0.0019}$**	0.4377$_{0.0032}$	0.6729$_{0.0143}$	0.6901$_{0.0135}$	0.7039$_{0.0298}$	2.2818	0.6779$_{0.1614}$	1.6877$_{0.5051}$
	SERA (↓)	46061	6564.0699$_{73.6302}$	6807.1688$_{175.8786}$	6538.4622$_{100.7926}$	*6528.1300$_{68.8213}$*	6549.7040$_{93.1419}$	**6362.0409$_{188.1432}$**	108670.1584	21496.1178$_{663.877}$	35728.3909$_{7867.882}$
		46076	8518.9647$_{44.5683}$	8435.2063$_{128.3628}$	8398.3464$_{185.7868}$	7435.9183$_{102.6282}$	*7293.7648$_{69.0909}$*	**7252.7616$_{125.3626}$**	119914.2962	35365.4742$_{1458.6353}$	190805.678$_{68279.1333}$
		46078	18394.7586$_{104.3305}$	18470.6403$_{233.2760}$	18455.2227$_{89.3421}$	**13675.3231$_{17.7498}$**	13705.4455$_{29.6155}$	*13687.1571$_{24.8219}$*	118435.5287	55901.172$_{25037.2841}$	347702.3309$_{70787.8264}$
		46082	10278.7888$_{194.9226}$	9867.1941$_{206.0941}$	9739.1154$_{152.4571}$	**7594.4811$_{17.5420}$**	7790.6611$_{113.7059}$	*7668.2098$_{104.0323}$*	102057.5456	25856.6713$_{9435.9262}$	58727.2112$_{18388.7018}$
		46085	15217.4397$_{95.4788}$	14935.6931$_{114.5193}$	15231.3725$_{298.5966}$	**11765.2626$_{82.1648}$**	11881.4101$_{154.1414}$	*11875.6736$_{144.0537}$*	179228.8871	38097.2518$_{11644.4851}$	128592.7468$_{35288.4399}$
	R^2 (↑)	46061	*0.8556$_{0.0013}$*	0.8548$_{0.0027}$	**0.8588$_{0.0013}$**	0.8452$_{0.0018}$	0.8499$_{0.0027}$	0.8515$_{0.0054}$	0.0000	0.7555$_{0.0024}$	0.7395$_{0.0114}$
		46076	0.8746$_{0.0003}$	*0.8756$_{0.0011}$*	**0.8757$_{0.0016}$**	0.8696$_{0.0031}$	0.8718$_{0.0031}$	0.8727$_{0.0057}$	0.0000	0.7446$_{0.0015}$	0.2587$_{0.2030}$
		46078	**0.7188$_{0.0003}$**	*0.7177$_{0.0016}$*	0.7184$_{0.0004}$	0.6879$_{0.0042}$	0.6820$_{0.0018}$	0.6866$_{0.0082}$	0.0000	0.5583$_{0.0201}$	0.1738$_{0.1352}$
		46082	0.7759$_{0.0015}$	*0.7760$_{0.0015}$*	**0.7790$_{0.0013}$**	0.7704$_{0.0036}$	0.7639$_{0.0082}$	0.7665$_{0.0071}$	0.0000	0.7265$_{0.0059}$	0.7198$_{0.0232}$
		46085	*0.7945$_{0.0006}$*	**0.7950$_{0.0008}$**	0.7934$_{0.0015}$	0.7883$_{0.0032}$	0.7827$_{0.0042}$	0.7854$_{0.0072}$	0.0000	0.7473$_{0.0076}$	0.6782$_{0.0369}$
	MaxSE (↓)	46061	5.5329$_{0.0753}$	5.1899$_{0.1370}$	5.3058$_{0.1358}$	5.0635$_{0.4525}$	**4.4701$_{0.2551}$**	*4.5472$_{0.3304}$*	29.5771	12.7718$_{0.3341}$	15.2489$_{1.1278}$
		46076	11.1380$_{0.1235}$	11.0649$_{0.3006}$	10.9349$_{0.3733}$	9.4982$_{0.1847}$	*8.9685$_{0.1928}$*	**8.8731$_{0.3735}$**	41.002	19.6955$_{0.7686}$	37.9584$_{9.4604}$
		46078	16.4247$_{0.1945}$	16.6703$_{0.6856}$	16.5127$_{0.1759}$	**12.7660$_{0.0811}$**	12.8762$_{0.2287}$	*12.8489$_{0.1391}$*	45.6979	33.1168$_{6.1929}$	75.8674$_{6.0879}$
		46082	7.2356$_{0.1366}$	7.1569$_{0.1879}$	7.0188$_{0.1963}$	*5.6095$_{0.1156}$*	5.7074$_{0.4792}$	**5.4034$_{0.3683}$**	27.3254	9.2766$_{2.2289}$	15.3163$_{3.2001}$
		46085	8.6043$_{0.0433}$	8.7744$_{0.2632}$	8.6100$_{0.1298}$	**6.4433$_{0.9217}$**	6.8300$_{0.1869}$	*6.6228$_{0.2882}$*	37.4499	15.1507$_{3.2015}$	34.9317$_{5.5485}$
	MRE (↓)	46061	0.1631$_{0.0008}$	*0.1621$_{0.0017}$*	**0.1599$_{0.0007}$**	0.1950$_{0.0062}$	0.1898$_{0.0076}$	0.1888$_{0.0078}$	0.4611	0.1973$_{0.0030}$	0.2606$_{0.0390}$
		46076	*0.1493$_{0.0005}$*	**0.1491$_{0.0007}$**	*0.1493$_{0.0008}$*	0.1673$_{0.0048}$	0.1741$_{0.0076}$	0.1803$_{0.0093}$	0.5034	0.2100$_{0.0064}$	0.4273$_{0.1068}$
		46078	**0.2024$_{0.0003}$**	0.2034$_{0.0012}$	*0.2029$_{0.0003}$*	0.3495$_{0.0030}$	0.3434$_{0.0018}$	0.3440$_{0.0044}$	0.4211	0.3506$_{0.1167}$	0.6547$_{0.1028}$
		46082	*0.1739$_{0.0006}$*	*0.1739$_{0.0006}$*	**0.1728$_{0.0007}$**	0.2196$_{0.0030}$	0.2216$_{0.0067}$	0.2258$_{0.0099}$	0.3968	0.2255$_{0.0474}$	0.2877$_{0.0547}$
		46085	**0.1637$_{0.0003}$**	0.1641$_{0.0005}$	*0.1640$_{0.0005}$*	0.2264$_{0.0040}$	0.2279$_{0.0041}$	0.2332$_{0.0078}$	0.3907	0.2118$_{0.0376}$	0.3238$_{0.0637}$
	Metric	Buoy	Optimised by MSE			Optimised by SERA			Mean	BRITS	SAITS
			PU	SU	PU+SU	PU	SU	PU+SU			
Northeast Coast	MSE (↓)	44005	*0.1053$_{0.0044}$*	**0.1031$_{0.0034}$**	0.1117$_{0.0110}$	0.1188$_{0.0086}$	0.1209$_{0.0081}$	0.1199$_{0.0098}$	0.9220	0.2981$_{0.0021}$	0.3193$_{0.0180}$
		44008	**0.1514$_{0.0041}$**	0.1563$_{0.0086}$	*0.1541$_{0.0058}$*	0.7886$_{2.6638}$	0.3378$_{0.0872}$	0.3215$_{0.0768}$	0.8217	0.2196$_{0.0586}$	0.1965$_{0.0193}$
		44011	0.8465$_{0.0056}$	0.8433$_{0.0232}$	*0.8380$_{0.0292}$*	1.4533$_{0.0365}$	1.5241$_{0.0886}$	1.5010$_{1.5010}$	1.8373	**0.6534$_{0.1993}$**	0.9191$_{0.2380}$
		44020	0.0804$_{0.0907}$	**0.0412$_{0.0007}$**	*0.0419$_{0.0008}$*	0.0758$_{0.0082}$	0.0729$_{0.0038}$	0.0744$_{0.0047}$	0.1053	0.0647$_{0.0006}$	0.0633$_{0.0059}$
		44025	0.0709$_{0.1759}$	*0.0332$_{0.0005}$*	**0.0330$_{0.0005}$**	0.0468$_{0.0121}$	0.0438$_{0.0027}$	0.0438$_{0.0025}$	0.5566	0.1348$_{0.0018}$	0.1341$_{0.0063}$
		44027	0.4560$_{0.0007}$	*0.4557$_{0.0012}$*	**0.4536$_{0.0021}$**	0.5685$_{0.0077}$	0.5643$_{0.0099}$	0.5659$_{0.0136}$	0.6934	0.5347$_{0.1575}$	0.5697$_{0.0889}$
		44065	*0.0333$_{0.0002}$*	**0.0330$_{0.0008}$**	**0.0330$_{0.0007}$**	0.0434$_{0.0015}$	0.0433$_{0.0019}$	0.0424$_{0.0027}$	0.2969	0.1235$_{0.0045}$	0.1925$_{0.0456}$
		44066	**0.1158$_{0.0024}$**	*0.1179$_{0.0030}$*	0.1186$_{0.0032}$	0.1648$_{0.0079}$	0.1859$_{0.0191}$	0.1723$_{0.0131}$	0.9407	0.2963$_{0.0081}$	0.3063$_{0.0195}$
	SERA (↓)	44005	3613.8875$_{271.4587}$	3744.3062$_{229.8895}$	3907.2206$_{574.5327}$	**3482.6442$_{218.9006}$**	3560.3486$_{244.3136}$	*3528.1412$_{370.5786}$*	76562.0017	18452.6337$_{298.8819}$	20063.3511$_{1585.0539}$
		44008	*4846.1251$_{137.6652}$*	5271.7589$_{206.6653}$	**4835.3828$_{231.8689}$**	5267.8654$_{2523.6104}$	5275.3799$_{485.7511}$	5113.9226$_{500.1604}$	25414.3219	10734.3772$_{3144.1024}$	12454.8032$_{1499.5253}$
		44011	16669.7033$_{473.6500}$	16868.4807$_{1423.8097}$	17009.1688$_{1311.5456}$	**11394.5672$_{130.4803}$**	13000.5643$_{2028.2205}$	*11752.9770$_{11752.9770}$*	105814.0418	19382.0014$_{3813.4157}$	45427.7612$_{12944.6109}$
		44020	**1259.9140$_{132.9363}$**	1796.7355$_{42.3087}$	1840.4305$_{69.4784}$	*1313.7564$_{38.3863}$*	1326.6046$_{27.1161}$	1319.3581$_{35.3638}$	11152.3129	5456.1263$_{181.0929}$	6267.9095$_{1135.2531}$
		44025	1261.1829$_{257.3111}$	1172.6459$_{70.9526}$	*1116.3864$_{46.8647}$*	1124.1819$_{29.6539}$	1127.3527$_{47.4385}$	**1099.3406$_{42.0522}$**	52985.298	11131.7309$_{250.2592}$	11246.6733$_{656.0384}$
		44027	25648.5797$_{106.1290}$	25569.0626$_{187.8732}$	25436.7969$_{239.5733}$	18035.6725$_{77.6232}$	*17908.8358$_{299.8570}$*	**17739.7083$_{350.6746}$**	59447.9833	40590.4238$_{17504.0232}$	60558.0768$_{10596.6004}$
		44065	1537.4463$_{18.9930}$	1519.2521$_{59.6646}$	1506.1100$_{50.6394}$	1431.9683$_{24.4679}$	**1408.6447$_{23.6190}$**	*1417.7112$_{27.6952}$*	29501.1561	9866.6322$_{611.0786}$	19196.6793$_{3995.2856}$
		44066	6118.0188$_{301.4930}$	6296.0088$_{273.4421}$	6252.8932$_{235.3900}$	*5616.5070$_{220.0157}$*	5680.5942$_{242.2960}$	**5584.9080$_{267.1034}$**	93020.5637	26960.5576$_{1914.169}$	30026.9692$_{2852.8357}$
	R^2 (↑)	44005	*0.8826$_{0.0041}$*	**0.8854$_{0.0040}$**	0.8767$_{0.0106}$	0.8732$_{0.0110}$	0.8721$_{0.0074}$	0.8733$_{0.0091}$	0.0000	0.6728$_{0.0007}$	0.6622$_{0.0138}$
		44008	**0.7754$_{0.0055}$**	0.7716$_{0.0125}$	*0.7727$_{0.0077}$*	0.6966$_{0.1342}$	0.7099$_{0.0544}$	0.7244$_{0.0539}$	0.0000	0.7274$_{0.0046}$	0.7242$_{0.0151}$
		44011	0.5142$_{0.0038}$	0.5183$_{0.0134}$	0.5159$_{0.0142}$	0.5264$_{0.0075}$	0.4709$_{0.0314}$	0.5145$_{0.5144}$	0.0000	**0.6893$_{0.0074}$**	*0.6784$_{0.0217}$*
		44020	**0.8788$_{0.1111}$**	*0.6105$_{0.0061}$*	0.6048$_{0.0074}$	0.5847$_{0.0444}$	0.6016$_{0.0091}$	0.5968$_{0.0122}$	0.0000	0.4200$_{0.0007}$	0.4564$_{0.0198}$
		44025	0.9119$_{0.1159}$	*0.9413$_{0.0009}$*	**0.9419$_{0.0009}$**	0.9318$_{0.0184}$	0.9369$_{0.0020}$	0.9387$_{0.0035}$	0.0000	0.7668$_{0.0008}$	0.7690$_{0.0073}$
		44027	0.3557$_{0.0006}$	0.3554$_{0.0009}$	0.3569$_{0.0011}$	0.3565$_{0.0025}$	0.3589$_{0.0029}$	0.3598$_{0.0027}$	0.0000	**0.3888$_{0.0088}$**	*0.3711$_{0.0164}$*
		44065	0.8881$_{0.0007}$	*0.8890$_{0.0028}$*	**0.8891$_{0.0023}$**	0.8870$_{0.0023}$	0.8868$_{0.0025}$	0.8869$_{0.0024}$	0.0000	0.6140$_{0.0053}$	0.6316$_{0.0106}$
		44066	**0.8788$_{0.0014}$**	*0.8764$_{0.0030}$*	0.8761$_{0.0036}$	0.8593$_{0.0114}$	0.8641$_{0.0126}$	0.8659$_{0.0089}$	0.0000	0.6933$_{0.0026}$	0.7016$_{0.0153}$
	MaxSE (↓)	44005	4.5062$_{1.5749}$	**3.8340$_{0.7367}$**	4.5098$_{1.2862}$	*4.3951$_{1.0163}$*	4.5183$_{0.9597}$	4.8000$_{1.6630}$	31.6402	16.6118$_{0.2639}$	18.854$_{1.6851}$
		44008	*4.2322$_{3.3110}$*	**3.9289$_{0.2078}$**	4.4385$_{3.8965}$	489.1272$_{2564.7203}$	5.7118$_{3.9921}$	7.1985$_{7.8524}$	9.2131	5.6618$_{1.0389}$	6.2085$_{0.3300}$
		44011	*12.8397$_{0.1075}$*	14.5254$_{3.4158}$	13.7603$_{3.4804}$	17.1146$_{1.1567}$	17.6500$_{7.4322}$	18.3466$_{18.3466}$	37.5022	**10.6181$_{2.0302}$**	15.3156$_{1.8099}$
		44020	67.8355$_{132.0093}$	4.3212$_{0.0786}$	4.3301$_{0.0906}$	6.9224$_{9.3477}$	3.7389$_{0.1769}$	*3.7366$_{0.2224}$*	**3.3333**	4.6248$_{0.0243}$	4.6229$_{0.1058}$
		44025	53.7501$_{256.3741}$	0.6409$_{0.2221}$	0.6774$_{0.4113}$	4.6355$_{16.0142}$	**0.5840$_{0.0962}$**	*0.6379$_{0.3755}$*	28.0251	8.5925$_{0.1543}$	8.9943$_{0.2550}$
		44027	19.3549$_{0.0728}$	19.1377$_{0.1745}$	19.2003$_{0.1412}$	15.2201$_{0.0638}$	14.9912$_{0.3099}$	*14.8523$_{0.3427}$*	25.5626	**12.8562$_{2.6673}$**	15.7778$_{1.7239}$
		44065	**1.0901$_{0.0285}$**	*1.1201$_{0.1149}$*	1.1451$_{0.0900}$	1.3645$_{0.0543}$	1.4717$_{0.0557}$	1.4855$_{0.0546}$	15.8524	6.3021$_{0.2637}$	6.7871$_{0.8034}$
		44066	6.9623$_{0.7734}$	7.7921$_{1.4113}$	7.6959$_{1.3778}$	**6.3533$_{0.7791}$**	7.0904$_{1.2825}$	*6.9607$_{2.2956}$*	57.0214	33.1009$_{1.2021}$	34.8268$_{1.7679}$
	MRE (↓)	44005	*0.1475$_{0.0027}$*	**0.1469$_{0.0020}$**	0.1515$_{0.0060}$	0.1679$_{0.0055}$	0.1703$_{0.0077}$	0.1699$_{0.0087}$	0.4271	0.2439$_{0.0006}$	0.2512$_{0.0072}$
		44008	**0.1841$_{0.0013}$**	*0.1852$_{0.0070}$*	0.1856$_{0.0039}$	0.2883$_{0.0289}$	0.3221$_{0.0553}$	0.3117$_{0.0371}$	0.5376	0.2431$_{0.0433}$	0.2048$_{0.0095}$
		44011	0.2827$_{0.0010}$	0.2814$_{0.0030}$	*0.2813$_{0.0032}$*	0.4054$_{0.0061}$	0.4153$_{0.0114}$	0.4140$_{0.0209}$	0.4082	**0.2567$_{0.0531}$**	0.2861$_{0.0419}$
		44020	0.2922$_{0.0030}$	**0.2897$_{0.0031}$**	*0.2915$_{0.0031}$*	0.4306$_{0.0144}$	0.4250$_{0.0173}$	0.4302$_{0.0215}$	0.4709	0.3474$_{0.0029}$	0.3384$_{0.0141}$
		44025	*0.1053$_{0.0037}$*	**0.1015$_{0.0007}$**	**0.1015$_{0.0007}$**	0.1254$_{0.0069}$	0.1251$_{0.0068}$	0.1261$_{0.0064}$	0.4251	0.1866$_{0.0051}$	0.1833$_{0.0048}$
		44027	0.3894$_{0.0007}$	**0.3883$_{0.0010}$**	*0.3884$_{0.0021}$*	0.4761$_{0.0068}$	0.4721$_{0.0114}$	0.4744$_{0.0147}$	0.5332	0.4512$_{0.0813}$	0.4187$_{0.0331}$
		44065	0.1255$_{0.0004}$	**0.1250$_{0.0015}$**	*0.1251$_{0.0012}$*	0.1514$_{0.0048}$	0.1506$_{0.0072}$	0.1476$_{0.0096}$	0.3729	0.2347$_{0.0099}$	0.2965$_{0.0486}$
		44066	**0.1485$_{0.0012}$**	*0.1492$_{0.0019}$*	0.1496$_{0.0015}$	0.1981$_{0.0073}$	0.2191$_{0.0149}$	0.2072$_{0.0133}$	0.4373	0.2266$_{0.0083}$	0.2199$_{0.0089}$

The best result is highlighted in **bold**; the second one best result is shown in *italics*.

Furthermore, it is worth of mention that our approach is not only better in terms of performance but also in terms of complexity. Table 12 compares the number of connections for each of the buoys of both coastal areas. The number of connections is expressed as their mean and Standard Deviation (*Mean$_{SD}$*). Nonetheless, BRITS and SAITS have the same number of connections regardless of the run. As can be observed, EANN models have a maximum of 27.40 connections on average, whereas the minimum number of connections for BRITS and SAITS models is 196 and 1319960, respectively. In addition, our approach does not require the use of a GPU, whereas BRITS or SAITS do, leading to heavier structures.

6. Conclusions

This paper presents a novel technique for the massive recovery of missing Significant Wave Height (SWH) time series data. The existence of missing data precludes the use of Machine Learning (ML) techniques for the estimation of future values, as they require a training set for building the models. Specifically, the proposed approach includes two phases for recovering missing data in a set of buoys. The first phase provides an intermediate recovery of each buoy, which will be used in the second phase to carry out the final reconstruction of the other buoys belonging to the same coastal zone but not for themselves. In this sense, three different Transfer Function (TF) models are presented for the first phase: regression-based, correlation-based and distance-based. From these three TFs, the distance-based TF stands out for achieving the best results, demonstrating that weighting by distance between the buoys of the same zone is capable of capturing the complex dynamics of the time series to be intermediately recovered. Once all buoys have been intermediately recovered, the final recovery is performed using correlated buoys as input. ML techniques can be applied for this second phase, as all input

Table 12
Comparison between the models of the second phase of the approach optimised by MSE and SERA, respectively, in terms of the number of connections and for each of the buoys of both coastal zones. The number of connections of the stochastic models is expressed as their mean and Standard Deviation (*SD*): Mean_{SD}.

	Buoys	Optimised by MSE			Optimised by SERA			BRITS	SAITS
		PU	SU	PU+SU	PU	SU	PU+SU		
Gulf of Alaska	46061	$\mathbf{6.97_{1.19}}$	$11.90_{0.84}$	$11.23_{0.63}$	$\mathit{10.23_{0.73}}$	$14.87_{0.63}$	$\mathit{10.23_{0.90}}$	284	1321514
	46076	$\mathbf{8.30_{1.70}}$	$12.43_{1.19}$	$12.67_{1.12}$	$12.30_{0.92}$	$16.07_{1.36}$	$\mathit{11.17_{1.37}}$	388	1323072
	46078	$\mathbf{3.33_{0.76}}$	$9.30_{1.51}$	$\mathit{8.87_{0.57}}$	$9.07_{0.37}$	$13.00_{0.00}$	$8.93_{0.37}$	196	1319960
	46082	$\mathbf{5.40_{1.16}}$	$12.23_{1.19}$	$12.13_{0.82}$	$\mathit{11.50_{1.04}}$	$16.67_{1.03}$	$11.73_{1.11}$	388	1323072
	46085	$\mathbf{6.80_{0.89}}$	$11.57_{1.14}$	$10.83_{1.37}$	$10.67_{0.61}$	$15.13_{0.68}$	$\mathit{9.70_{1.02}}$	284	1321514
	Buoys	Optimised by MSE			Optimised by SERA			BRITS	SAITS
		PU	SU	PU+SU	PU	SU	PU+SU		
Northeast Coast	44005	$\mathbf{6.77_{1.89}}$	$27.33_{2.66}$	$\mathit{10.93_{2.33}}$	$15.90_{2.16}$	$20.30_{2.09}$	$19.43_{1.98}$	796	1327770
	44008	$\mathbf{5.57_{1.43}}$	$15.03_{1.90}$	$\mathit{13.87_{2.50}}$	$16.07_{2.13}$	$21.80_{2.12}$	$19.90_{2.70}$	964	1329344
	44011	$\mathbf{8.63_{1.22}}$	$19.60_{1.61}$	$\mathit{10.57_{1.57}}$	$11.40_{1.04}$	$16.73_{1.08}$	$14.13_{14.13}$	388	1323072
	44020	$\mathbf{10.67_{1.58}}$	$16.97_{1.83}$	$14.83_{1.97}$	$\mathit{13.83_{1.46}}$	$19.83_{1.66}$	$27.40_{1.75}$	644	1326200
	44025	$\mathbf{7.30_{1.70}}$	$24.37_{1.87}$	$16.57_{1.59}$	$\mathit{12.27_{1.20}}$	$17.77_{1.38}$	$16.03_{1.83}$	508	1324634
	44027	$\mathbf{7.37_{1.71}}$	$\mathit{10.97_{0.93}}$	$16.60_{2.76}$	$16.33_{1.15}$	$23.83_{1.15}$	$15.13_{1.43}$	388	1323072
	44065	$\mathbf{6.33_{1.32}}$	$11.47_{0.78}$	$10.60_{0.62}$	$\mathit{10.43_{0.68}}$	$14.83_{0.70}$	$13.63_{0.61}$	284	1321514
	44066	$\mathbf{8.63_{1.90}}$	$13.90_{1.88}$	$\mathit{11.77_{2.93}}$	$15.83_{1.60}$	$20.90_{1.77}$	$19.07_{2.23}$	796	1327770

The lowest number of connections is highlighted in **bold**; the second one lowest is shown in *italics*.

buoys have available values at the same time instants, either originally collected or intermediately recovered. To tackle this second phase, Evolutionary Artificial Neural Networks (EANNs) have been proposed using Product Units (PUs), Sigmoid Units (SUs) and hybrid networks combining PU with SUs (PU+SUs) as basis functions for the neurons in the hidden layer. Moreover, a comprehensive comparison of ML techniques has been carried out by means of 7 state-of-the-art techniques: Multilayer Perceptron, Support Vector Machines or Random Forest, among others. These techniques have been trained using two different metrics, each promoting a different strategy. On the one hand, the Mean Squared Error (MSE) tends to follow the general trend of the SWH time series. However, extreme values of SWH time series are underestimated. On the other hand, a novel metric known as Squared Error-Relevance Area (SERA) has also been used, which aims to balance between common values (around the mean) and extreme values (which can even be considered outliers). The proposed approach has been applied to two coastal areas of the United States of America: the Gulf of Alaska and the Northeast Coast, where 6 and 9 buoys have been considered, respectively. With respect to the results obtained, EANNs demonstrated their superiority over the state-of-the-art techniques, with PU and PU+SU being the top performers. Moreover, the results demonstrate that both metrics are excellent in achieving the objectives they have been designed for. The end user will be able to decide which metric is the most appropriate depending on the specific situation to be solved.

Ultimately, this proposal can be extended in four directions as future lines of work. Firstly, the proposed methodology can be applied to a wide variety of areas where data is collected by sensors or devices located geographically close. For example, this approach could be immediately applied to wind or solar energy farms in the green energies field or even to fog and rainfall detection systems at airports in the atmospheric area. Secondly, the use of exogenous variables to the SWH time series, either collected from the buoys themselves or mathematical reanalysis models, could improve the results. Thirdly, to carefully study whether the SERA and MSE metrics are opposing, i.e. optimising one implies a decrease in performance in the other metric. If so, a multi-objective algorithm may be an excellent approach to enhance the proposed methodology. Finally, we believe that the main limitation of the proposed methodology concerns the evaluation of EANNs, as it is the most computationally expensive step. Therefore, we propose to reduce the computational complexity of our approach by parallelising such evaluation.

CRediT authorship contribution statement

David Guijo-Rubio: Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Writing – review & editing, Visualization. **Antonio M. Durán-Rosal:** Methodology, Software, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Antonio M. Gómez-Orellana:** Software, Investigation, Validation, Visualization, Conceptualization, Writing – original draft, Writing – review & editing. **Juan C. Fernández:** Software, Investigation, Validation, Resources, Supervision, Project administration, 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.

Data availability

Data used is freely available on the NDBC and NOAA website.

Acknowledgements

This work was supported by the "Agencia Estatal de Investigación (España)" (grant reference: PID2020-115454GB-C22 / AEI / 10.13039 / 501100011033); the "Consejería de Salud y Familia (Junta de Andalucía)" (grant reference: PS-2020-780); and the "Consejería de Transformación Económica, Industria, Conocimiento y Universidades (Junta de Andalucía) y Programa Operativo FEDER 2014–2020" (grant references: UCO-1261651 and PY20_00074). David Guijo-Rubio's research has been subsidised by the University of Córdoba through grants to Public Universities for the requalification of the Spanish university system of the Ministry of Universities, financed by the European Union - NextGenerationEU (grant reference: UCOR01MS). Antonio M. Gómez-Orellana's research has been supported by "Consejería de Transformación Económica, Industria, Conocimiento y Universidades de la Junta de Andalucía" (grant reference: PREDOC-00489). Funding for open access charge: Universidad de Córdoba/CBUA.

References

[1] M. Katalinić, J. Parunov, Uncertainties of estimating extreme significant wave height for engineering applications depending on the approach and fitting technique—Adriatic sea case study, J. Mar. Sci. Eng. 8 (4) (2020) 259, http://dx.doi.org/10.3390/jmse8040259.

[2] C. Chen, Case study on wave-current interaction and its effects on ship navigation, J. Hydrodyn. 30 (3) (2018) 411–419, http://dx.doi.org/10.1007/s42241-018-0050-5.

[3] X. Yang, R. Ramezani, I.B. Utne, A. Mosleh, P.F. Lader, Operational limits for aquaculture operations from a risk and safety perspective, Reliab. Eng. Syst. Saf. 204 (2020) 107208, http://dx.doi.org/10.1016/j.ress.2020.107208.

[4] A. Masoumi, S. Ghassem-zadeh, S.H. Hosseini, B.Z. Ghavidel, Application of neural network and weighted improved PSO for uncertainty modeling and optimal allocating of renewable energies along with battery energy storage, Appl. Soft Comput. 88 (2020) 105979, http://dx.doi.org/10.1016/j.asoc.2019.105979.

[5] NDBC, National data buoy center, national oceanic and atmospheric administration of the USA (NOAA), 2022, URL https://www.ndbc.noaa.gov/wavecalc.shtml. (Accessed 2 September 2022).

[6] C. Rodrigues, M. Ramos, R. Esteves, J. Correia, D. Clemente, F. Gonçalves, N. Mathias, M. Gomes, J. Silva, C. Duarte, et al., Integrated study of triboelectric nanogenerator for ocean wave energy harvesting: Performance assessment in realistic sea conditions, Nano Energy 84 (2021) 105890, http://dx.doi.org/10.1016/j.nanoen.2021.105890.

[7] F. Barbariol, A. Benetazzo, L. Bertotti, L. Cavaleri, T. Durrant, P. McComb, M. Sclavo, Large waves and drifting buoys in the Southern Ocean, Ocean Eng. 172 (2019) 817–828, http://dx.doi.org/10.1016/j.oceaneng.2018.12.011.

[8] A.S. Kuppili, Forecasting Meteorological Variables and Anticipating Climatic Aberrations of an Oceanic Buoy Using A Neighbour Buoy (Ph.D. thesis), Dalhousie University, 2021.

[9] M. Jeon, Y. Noh, K. Jeon, S. Lee, I. Lee, Data gap analysis of ship and maritime data using meta learning, Appl. Soft Comput. 101 (2021) 107048, http://dx.doi.org/10.1016/j.asoc.2020.107048.

[10] L. Carro-Calvo, F. Jaume-Santero, R. García-Herrera, S. Salcedo-Sanz, K-gaps: a novel technique for clustering incomplete climatological time series, Theor. Appl. Climatol. 143 (1) (2021) 447–460, http://dx.doi.org/10.1007/s00704-020-03396-w.

[11] L.H. Chua, T.S. Wong, X. Wang, Information recovery from measured data by linear artificial neural networks—An example from rainfall–runoff modeling, Appl. Soft Comput. 11 (1) (2011) 373–381, http://dx.doi.org/10.1016/j.asoc.2009.11.028.

[12] E. Visek, L. Mazzarella, M. Motta, Temperature sensor signal reconstruction for failure detection of vapor compression system, Appl. Soft Comput. 60 (2017) 679–688, http://dx.doi.org/10.1016/j.asoc.2017.06.054.

[13] J. Chen, F. Zhu, Y. Han, Z. Xu, Q. Chen, D. Ren, Global temperature reconstruction of equipment based on the local temperature image using TRe-GAN, Appl. Soft Comput. (2022) 109498, http://dx.doi.org/10.1016/j.asoc.2022.109498.

[14] M. Jović, E. Tijan, R. Marx, B. Gebhard, Big data management in maritime transport, Pomorski zbornik 57 (1) (2019) 123–141, http://dx.doi.org/10.18048/2019.57.09.

[15] S. Yusop, M. Mustapha, Influence of oceanographic parameters on the seasonal potential fishing grounds of rastrelliger kanagurta using maximum entropy models and remotely sensed data, Sains Malaysiana 48 (2) (2019) 259–269, http://dx.doi.org/10.17576/jsm-2019-4802-01.

[16] S. Emmanouil, S.G. Aguilar, G.F. Nane, J.-J. Schouten, Statistical models for improving significant wave height predictions in offshore operations, Ocean Eng. 206 (2020) 107249, http://dx.doi.org/10.1016/j.oceaneng.2020.107249.

[17] E. Vanem, Joint statistical models for significant wave height and wave period in a changing climate, Mar. Struct. 49 (2016) 180–205, http://dx.doi.org/10.1016/j.marstruc.2016.06.001.

[18] F. Wang, D. Yang, L. Yang, Retrieval and assessment of significant wave height from CYGNSS mission using neural network, Remote Sens. 14 (15) (2022) 3666, http://dx.doi.org/10.3390/rs14153666.

[19] C.W. Zheng, C.Y. Li, Variation of the wave energy and significant wave height in the China Sea and adjacent waters, Renew. Sustain. Energy Rev. 43 (2015) 381–387, http://dx.doi.org/10.1016/j.rser.2014.11.001.

[20] Y. Lin, S. Dong, Wave energy assessment based on trivariate distribution of significant wave height, mean period and direction, Appl. Ocean Res. 87 (2019) 47–63, http://dx.doi.org/10.1016/j.apor.2019.03.017.

[21] T. Caloiero, F. Aristodemo, D. Algieri Ferraro, Trend analysis of significant wave height and energy period in southern Italy, Theor. Appl. Climatol. 138 (1) (2019) 917–930, http://dx.doi.org/10.1007/s00704-019-02879-9.

[22] N. Guillou, Estimating wave energy flux from significant wave height and peak period, Renew. Energy 155 (2020) 1383–1393, http://dx.doi.org/10.1016/j.renene.2020.03.124.

[23] A.M. Gómez-Orellana, J.C. Fernández, M. Dorado-Moreno, P.A. Gutiérrez, C. Hervás-Martínez, 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) (2021) 468, http://dx.doi.org/10.3390/en14020468.

[24] L. Cornejo-Bueno, J. Nieto-Borge, P. García-Díaz, G. Rodríguez, S. Salcedo-Sanz, Significant wave height and energy flux prediction for marine energy applications: A grouping genetic algorithm–Extreme learning machine approach, Renew. Energy 97 (2016) 380–389, http://dx.doi.org/10.1016/j.renene.2016.05.094.

[25] F. Taveira-Pinto, P. Rosa-Santos, T. Fazeres-Ferradosa, Marine renewable energy, Renew. Energy 150 (2020) 1160–1164, http://dx.doi.org/10.1016/j.renene.2019.10.014.

[26] A.M. Durán-Rosal, C. Hervás-Martínez, A.J. Tallón-Ballesteros, A. Martínez-Estudillo, S. Salcedo-Sanz, Massive missing data reconstruction in ocean buoys with evolutionary product unit neural networks, Ocean Eng. 117 (2016) 292–301, http://dx.doi.org/10.1016/j.oceaneng.2016.03.053.

[27] R.P. Ribeiro, N. Moniz, Imbalanced regression and extreme value prediction, Mach. Learn. 109 (9) (2020) 1803–1835, http://dx.doi.org/10.1007/s10994-020-05900-9.

[28] R.O. Thompson, Spectral estimation from irregularly spaced data, IEEE Trans. Geosci. Electron. 9 (2) (1971) 107–110, http://dx.doi.org/10.1109/tge.1971.271476.

[29] W. Sturges, On interpolating gappy records for time-series analysis, J. Geophys. Res.: Oceans 88 (C14) (1983) 9736–9740, http://dx.doi.org/10.1029/jc088ic14p09736.

[30] C. Cunha, C.G. Soares, On the choice of data transformation for modelling time series of significant wave height, Ocean Eng. 26 (6) (1999) 489–506, http://dx.doi.org/10.1016/s0029-8018(98)00014-6.

[31] C.N. Stefanakos, G. Athanassoulis, A unified methodology for the analysis, completion and simulation of nonstationary time series with missing values, with application to wave data, Appl. Ocean Res. 23 (4) (2001) 207–220, http://dx.doi.org/10.1016/s0141-1187(01)00017-7.

[32] M. Lázaro-Gredilla, S. Van Vaerenbergh, N.D. Lawrence, Overlapping mixtures of Gaussian processes for the data association problem, Pattern Recognit. 45 (4) (2012) 1386–1395, http://dx.doi.org/10.1016/j.patcog.2011.10.004.

[33] S. Salcedo-Sanz, J.L. Rojo-Álvarez, M. Martínez-Ramón, G. Camps-Valls, Support vector machines in engineering: an overview, WIREs Data Min. Knowl. Discov. 4 (3) (2014) 234–267, http://dx.doi.org/10.1002/widm.1125.

[34] J. Del Ser, D. Casillas-Perez, L. Cornejo-Bueno, L. Prieto-Godino, J. Sanz-Justo, C. Casanova-Mateo, S. Salcedo-Sanz, Randomization-based machine learning in renewable energy prediction problems: Critical literature review, new results and perspectives, Appl. Soft Comput. 118 (2022) 108526, http://dx.doi.org/10.1016/j.asoc.2022.108526.

[35] C.M. Bishop, et al., Neural Networks for Pattern Recognition, Oxford University Press, 1995.

[36] S. Puca, B. Tirozzi, G. Arena, S. Corsini, et al., Neural network approach to the problem of recovering lost data in a network of marine buoys, in: The Eleventh International Offshore and Polar Engineering Conference, OnePetro, 2001, pp. 620–623.

[37] E.-L. Silva-Ramírez, R. Pino-Mejías, M. López-Coello, Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns, Appl. Soft Comput. 29 (2015) 65–74, http://dx.doi.org/10.1016/j.asoc.2014.09.052.

[38] C.E. Balas, L. Koç, L. Balas, Predictions of missing wave data by recurrent neuronets, J. Waterw. Port Coast. Ocean Eng. 130 (5) (2004) 256–265, http://dx.doi.org/10.1061/(asce)0733-950x(2004)130:5(256).

[39] O. Makarynskyy, D. Makarynska, Wave prediction and data supplementation with artificial neural networks, J. Coast. Res. 23 (4) (2007) 951–960, http://dx.doi.org/10.2112/04-0407.1.

[40] J. Wang, L. Aouf, A. Dalphinet, Y. Zhang, Y. Xu, D. Hauser, J. Liu, The wide swath significant wave height: An innovative reconstruction of significant wave heights from CFOSAT's SWIM and scatterometer using deep learning, Geophys. Res. Lett. 48 (6) (2021) e2020GL091276, http://dx.doi.org/10.1029/2020gl091276.

[41] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780, http://dx.doi.org/10.1162/neco.1997.9.8.1735.

[42] E.-J. Lee, J.-Y. Chae, J.-H. Park, Reconstruction of sea level data around the Korean Coast using Artificial neural network methods, J. Coast. Res. 95 (SI) (2020) 1172–1176, http://dx.doi.org/10.2112/si95-227.1.

[43] C. Jörges, C. Berkenbrink, B. Stumpe, Prediction and reconstruction of ocean wave heights based on bathymetric data using LSTM neural networks, Ocean Eng. 232 (2021) 109046, http://dx.doi.org/10.1016/j.oceaneng.2021.109046.

[44] M. Pirhooshyaran, L.V. Snyder, Forecasting, hindcasting and feature selection of ocean waves via recurrent and sequence-to-sequence networks, Ocean Eng. 207 (2020) 107424, http://dx.doi.org/10.1016/j.oceaneng.2020.107424.

[45] A. Petrowski, S. Ben-Hamida, Evolutionary Algorithms, in: Computer Engineering: Metaheuristics, Wiley, 2017, http://dx.doi.org/10.1002/9781119136378.

[46] J.D. Ser, E. Osaba, D. Molina, X.-S. Yang, S. Salcedo-Sanz, D. Camacho, S. Das, P.N. Suganthan, C.A.C. Coello, F. Herrera, Bio-inspired computation: Where we stand and what's next, Swarm Evol. Comput. 48 (2019) 220–250, http://dx.doi.org/10.1016/j.swevo.2019.04.008.

[47] E. Alexandre, L. Cuadra, J. Nieto-Borge, G. Candil-García, M. Del Pino, S. Salcedo-Sanz, A hybrid genetic algorithm—extreme learning machine approach for accurate significant wave height reconstruction, Ocean Model. 92 (2015) 115–123, http://dx.doi.org/10.1016/j.ocemod.2015.06.010.

[48] D.I. Gopinath, G. Dwarakish, Real-time prediction of waves using neural networks trained by particle swarm optimization, Int. J. Ocean Clim. Syst. 7 (2) (2016) 70–79, http://dx.doi.org/10.1177/1759313116642896.

[49] W. Wang, R. Tang, C. Li, P. Liu, L. Luo, A BP neural network model optimized by mind evolutionary algorithm for predicting the ocean wave heights, Ocean Eng. 162 (2018) 98–107, http://dx.doi.org/10.1016/j.oceaneng.2018.04.039.

[50] A. Gómez-Orellana, D. Guijo-Rubio, P. Gutiérrez, C. Hervás-Martínez, Simultaneous short-term significant wave height and energy flux prediction using zonal multi-task evolutionary artificial neural networks, Renew. Energy 184 (2022) 975–989, http://dx.doi.org/10.1016/j.renene.2021.11.122.

[51] National Data Buoy Center, National oceanic and atmospheric administration of the USA (NOAA), 2022, http://www.ndbc.noaa.gov/. (Accessed 2 September 2022).

[52] NDBC, How are significant wave height, dominant period, average period, and wave steepness calculated?, 2022, URL https://www.ndbc.noaa.gov/wavecalc.shtml. (Accessed 2 September 2022).

[53] NDBC, Measurement descriptions and units, 2022, URL https://www.ndbc.noaa.gov/measdes.shtml. (Accessed 2 September 2022).

[54] M.J. de Smith, M.F. Goodchild, P.A. Longley, Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools, third ed., Matador, 2009, p. 516.

[55] F. Fernández-Navarro, M. Carbonero-Ruz, D.B. Alonso, M. Torres-Jimenez, Global sensitivity estimates for neural network classifiers, IEEE Trans. Neural Netw. Learn. Syst. 28 (11) (2017) 2592–2604, http://dx.doi.org/10.1109/TNNLS.2016.2598657.

[56] F. Martínez-Estudillo, C. Hervás-Martínez, P. Gutiérrez, A. Martínez-Estudillo, Evolutionary product-unit neural networks classifiers, Neurocomputing 72 (1) (2008) 548–561, http://dx.doi.org/10.1016/j.neucom.2007.11.019, Machine Learning for Signal Processing (MLSP 2006) / Life System Modelling, Simulation, and Bio-inspired Computing (LSMS 2007).

[57] S. Ding, H. Li, C. Su, J. Yu, F. Jin, Evolutionary artificial neural networks: a review, Artif. Intell. Rev. (99) (2013) 251–260, http://dx.doi.org/10.1007/s10462-011-9270-6.

[58] A. Martínez-Estudillo, F. Martínez-Estudillo, C. Hervás-Martínez, N. García-Pedrajas, Evolutionary product unit based neural networks for regression, Neural Netw. 19 (4) (2006) 477–486, http://dx.doi.org/10.1016/j.neunet.2005.11.001.

[59] A. Martinez-Estudillo, C. Hervas-Martinez, F. Martinez-Estudillo, N. Garcia-Pedrajas, Hybridization of evolutionary algorithms and local search by means of a clustering method, IEEE Trans. Syst. Man Cybern. B 36 (3) (2006) 534–545, http://dx.doi.org/10.1109/TSMCB.2005.860138.

[60] P.A. Gutiérrez, C. Hervás, M. Carbonero, J.C. Fernández, Combined projection and kernel basis functions for classification in evolutionary neural networks, Neurocomputing 72 (13–15) (2009) 2731–2742, http://dx.doi.org/10.1016/j.neucom.2008.09.020.

[61] J. Wilson, S. Joye, Research Methods and Statistics: An Integrated Approach, SAGE Publications, 2016.

[62] M. Hubert, E. Vandervieren, An adjusted boxplot for skewed distributions, Comput. Statist. Data Anal. 52 (12) (2008) 5186–5201, http://dx.doi.org/10.1016/j.csda.2007.11.008.

[63] G. Brys, M. Hubert, A. Struyf, A robust measure of skewness, J. Comput. Graph. Statist. 13 (4) (2004) 996–1017, http://dx.doi.org/10.1198/106186004x12632.

[64] G.A. Seber, A.J. Lee, Linear Regression Analysis, John Wiley & Sons, 2012, http://dx.doi.org/10.1002/9780471722199.

[65] M. Gruber, Improving Efficiency By Shrinkage: The James–Stein and Ridge Regression Estimators, CRC Press, 2017.

[66] T. Park, G. Casella, The bayesian lasso, J. Amer. Statist. Assoc. 103 (482) (2008) 681–686, http://dx.doi.org/10.1198/016214508000000337.

[67] C. De Mol, E. De Vito, L. Rosasco, Elastic-net regularization in learning theory, J. Complexity 25 (2) (2009) 201–230, http://dx.doi.org/10.1016/j.jco.2009.01.002.

[68] I. Steinwart, A. Christmann, Support Vector Machines, Springer Science & Business Media, 2008, http://dx.doi.org/10.1007/978-0-387-77242-4.

[69] R. Vang-Mata, Multilayer Perceptrons: Theory and Applications, in: Computer Science, Technology and Applications Series, Nova Science Publishers, 2020.

[70] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32, http://dx.doi.org/10.1023/A:1010933404324.

[71] W. Cao, D. Wang, J. Li, H. Zhou, L. Li, Y. Li, BRITS: Bidirectional recurrent imputation for time series, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Inc., 2018.

[72] W. Du, D. Côté, Y. Liu, SAITS: Self-attention-based imputation for time series, Expert Syst. Appl. (2023) 119619, http://dx.doi.org/10.1016/j.eswa.2023.119619.

[73] J. Yoon, W.R. Zame, M. van der Schaar, Multi-directional recurrent neural networks: A novel method for estimating missing data, in: Time Series Workshop in International Conference on Machine Learning, 2017.

[74] Z. Che, S. Purushotham, K. Cho, D. Sontag, Y. Liu, Recurrent neural networks for multivariate time series with missing values, Sci. Rep. 8 (1) (2018) 6085, http://dx.doi.org/10.1038/s41598-018-24271-9.

[75] V. Fortuin, D. Baranchuk, G. Rätsch, S. Mandt, Gp-vae: Deep probabilistic time series imputation, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1651–1661.