https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Ocean Science
Discussions
Open Access
EGU

An EMD-PSO-LSSVM hybrid model for significant wave height prediction

Gang Tang [1], Haohao Du [1], Xiong Hu [1] , Yide Wang [2], Christophe Claramunt [3] and Shaoyang Men [4,*]

1.Logistics Engineering College, Shanghai Maritime University, Shanghai 201306, China

2.Institut d'Électronique et des Technologies du numérique (IETR), UMR CNTS 6164, Polytech Nantes-Site de la Chantrerie, 44306 Nantes, France

3.Naval Academy Research Institute, F-29240 Lanvéoc, France

4.School of Medical Information Engineering, Guangzhou University of Chinese Medicine, Guangzhou 510006, China

*** Correspondence:** Shaoyang Men (shaoyang.men@gzucm.edu.cn)

Abstract. Accurate and significant wave height prediction with a couple of hours of warning time should offer major safety improvements for coastal and ocean engineering applications. However, significant wave height phenomenon is nonlinear and nonstationary, which makes any prediction simulation a non-straightforward task. The aim of the research presented in this paper is to improve predicted significant wave height via a hybrid algorithm. Firstly, empirical mode decomposition (EMD) is used to preprocess the nonlinear data, which are decomposed into several simple signals. Then, least square support vector machine (LSSVM) with nonlinear learning ability is used to predict the significant wave height, and particle swarm optimization (PSO) is implemented to automatically perform the parameter selection in LSSVM modeling. The EMD-PSO-LSSVM model is used to predict the significant wave height for 1, 3 and 6 hours leading times of two stations in the offshore and deep-sea areas of the North Atlantic Ocean. The results show that the EMD-PSO-LSSVM model can remove the lag in the prediction timing of the single prediction models. Furthermore, the prediction accuracy of the EMD-LSSVM model that has not been optimized in the deep-sea area has been greatly improved, an improvement of the prediction accuracy of Coefficient of determination ($\boldsymbol{R^2}$) from 0.991, 0.982 and 0.959 to 0.993, 0.987 and 0.965, respectively, has been observed. The proposed new hybrid model shows good accuracy and provides an effective way to predict the significant wave height for the deep-sea area.

1 Introduction

Significant wave height prediction has many vital applications. For instance, it can improve the efficiency and safety of operations in marine and offshore environments (Duan et al., 2016a). Installation of offshore wind turbines, cargo transfer between ships, sea rescue and lifting and landing of helicopters or aircraft are other significant examples to mention (Richter et al., 2017). More precisely, an accurate estimation of the significant wave height is relevant to characterize the wave energy production from Wave Energy Converters (WECs) facilities (Cornejo-Bueno et al., 2018). Prediction information can help to provide motion compensation, which may prevent the crash of cargo in cargo transfer, improve the firing accuracy of ship-borne weapon systems, and performance of the motion control systems (Ra and Whang, 2006).

Over the past few years, several numerical methods have been developed to predict significant wave height, using either classical statistical methods, the artificially intelligent techniques based on linear and nonlinear models, or hybrid models (Hwang, 2006;Casas-Prat et al., 2014;Janssen, 2008). However, accurate prediction of significant wave height requires a large amount of sensor-based data while the computational complexity of the calculations is still relatively high and requires high-performance computers. Last but not least wave height predictions are still not always very accurate (Yoon et al., 2011;Browne et al., 2007). With the development of machine learning, time series analysis provides an easy and computationally efficient solution that will be mainly based on historical wave height data. Such modelling approaches will have the advantage of being relatively simple as based on previous data and wave patterns, this avoiding a lot of computational costs.

Early research studies on wave prediction using machine learning employed classical time series models, such as the auto-regressive (AR) model, auto-regressive moving average (ARMA) model, an autoregressive integrated moving average (ARIMA) model. Soares et al. (1996) applied AR models to describe time series of significant wave heights in two Portuguese coast locations. Later, AR models have been further generalized from the application of univariate models of the long-term time series of significant wave height to the case of the bivariate series of significant wave height and mean period (Guedes Soares and Cunha, 2000). However, the prediction based on a single AR model in harsh conditions and large prediction leading time fails to satisfy the expectations. To further improve the prediction performance, Agrawal and Deo (2002) adopted ARMA and ARIMA models to predict the wave height for 3, 6, 12, and 24 hours of offshore location in India. Despite the high efficiency and adaptiveness of classical time series models, prediction results in severe sea conditions are far from being accurate enough. Since waves are always nonstationary, this conflicts with the linear and stationary classical time series models' assumptions. Overall, these approaches are not suitable for predicting nonlinear and nonstationary waves.

In order to address the nonlinear component of ocean waves, intelligent-technique-based nonlinear models such as artificial neural networks (ANNs) models have been extensively studied. Such methods can carry out nonlinear simulations without a deep

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

understanding of the relationships between the input and output variables. Deo and Sridhar Naidu (1998) were amongst the first to apply an ANN applied exclusively as input parameters for real-time wave forecasting in the next 3-24 hours using wave characteristics, and compared the training results of different algorithms. In order to estimate large wave height and average wave periods, Deo et al. (2001) used wind velocity data and fetched data. Tsai et al. (2002) applied an ANN based on data from three wave graph stations in areas with different physical characteristics for short-term estimation of the wave height. Makarynskyy (2004) used ANN for the substantial wave height and for the subsequent 1-24 h forecast times, and to amend the forecasts by employing two different approaches. Mandal and Prabaharan (2006) used recurrent neural network (RNN) for wave height prediction in Marmugao, west coast of India, they concluded that the wave prediction using RNN is better than previous applications of neural networks. One of the limitations of the neural network approach is that it needs to find network parameters such as the number of hidden layers and neurons by trial and error, this being time consuming. Mahjoobi and Adeli Mosabbeb (2009) applied the SVM to predict the wave height, the analysis indicated that the SVM model had a reasonable precision, and compared to an ANN model, it took less computing time. Different experiments on the prediction effects in Lake Superior were carried out by Etemad-Shahidi and Mahjoobi (2009), and they compared the model trees and feedforward backpropagation ANNs. Their findings revealed that the model tree system was the most precise. Dixit et al. (2015) found the phenomenon of prediction time lag while using ANNs to predict ocean wave height. They used a discrete wavelet transform to enhance the predicted values and removed the lag in the prediction timing. Akbarifard and Radmanesh (2018) introduced an symbiotic organisms search (SOS) algorithm to predict the ocean wave heights. The findings showed that the SOS algorithm's performance performed better than that of the support vector regression, ANN, and simulating waves nearshore dynamic models. Fan et al. (2020) proposed a long short-term network for the quick prediction of significant wave height with higher accuracy than the conventional neural networks.

Significant wave height is a complicated, nonlinear, dynamic system, and it is impacted by various components (Valamanesh et al., 2016). The time series prediction of non-stationary data by using the ANN method will lead to the homogenization of the different characteristics of the original input data, which could affect the prediction accuracy. Accordingly, the non-stationarity of the time series of significant wave height and input variables should be reduced. To handle nonstationary features, the inputs for the corresponding data-driven models are need to be appropriately preprocessed. Hybrid models that combine preprocessing techniques with single prediction models are alternatives for more effective modeling. The wavelet analysis is a useful tool that can be used for nonstationary data (Rhif et al., 2019). Deka and Prahlada (2012) developed a wavelet neural network model in their study by hybridizing ANN with a wavelet transform, and the prediction results suggested that the hybrid models outperformed single models. Kaloop et al. (2020) designed the wavelet-PSO-ELM (WPSO-ELM) model for estimating the wave height belongs to coastal and deep-sea stations. The results demonstrated that the WPSO-ELM outperforms other models to predict the wave height in both hourly and daily leading times. Essentially, a linear and nonstationary solution is based on wavelet transform. It represents a signal through a linear combination of functions of the wavelet base. For nonlinear data, therefore, it may not be suitable (Huang and Wu, 2008). The other issue with wavelets is that they require defining a well-suited mother wavelet transform a priori (Chen et al., 2012). It is still an unresolved issue and generally requires a lengthy trial and error process (Prasad et al., 2017). In hybrid prediction models, a more effective decomposition technique is needed to overcome nonlinearity and non-stationarity instantaneously.

In the study of nonlinear and nonstationary datasets, a data-driven methodology known as empirical mode decomposition (EMD) is efficient and adaptive (Huang et al., 1998). The EMD multiresolution utility offers self-adaptability by avoiding the need for any basis function and mother wavelets. It functions as a dyadic filter that divides a large frequency band complex signal into relatively essential, time-scale components (Flandrin et al., 2004). Duan et al. (2016b) proposed EMD-SVR for the short-term prediction of ocean waves. The result indicated the EMD-SVR model shows good model performances and provides an effective way for the short-term prediction of nonlinear and nonstationary waves.

Based on these recent findings, the research presented in this paper integrated the EMD-PSO with the LSSVM models in order to improve prediction models' accuracy. LSSVM with nonlinear learning ability can be used for predictionEMD is an empirical analysis tool used for processing nonlinear and nonstationary datasets. Preprocessing with EMD can reduce the difficulty of prediction. and PSO is a swarm intelligence optimization algorithm by updating the distance between the current and best locations. The important parameters of LSSVM are optimally adjusted by PSO to improve the prediction accuracy of a single LSSVM.

2 Methodology formulation

2.1 EMD-PSO-LSSVM prediction model

Ocean wave time series is a kind of complicated nonlinear and nonstationary signal composed of various oscillation scales. When performing wave predictions, the different oscillation scales create difficulties for the LSSVM models. Integrating an EMD model with an LSSVM model is an important way for enhancing the wave hright prediciton.. The EMD was adopted to decompose the wave height series that consisted of one residual series and several intrinsic mode functions (IMFs). Then, the residual series and IMFs were modeled by the LSSVM model, respectively, and finally the summation of the prediction outputs of subseries the wave height prediction. Besides, the PSO was employed to optimize LSSVM parameters to increase prediction accuracy. The specific steps of the EMD-PSO-LSSVM prediction algorithm are displayed in Fig. 1. The next part is to present this hybrid technique separately.

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Ocean Science
Discussions
Open Access
EGU

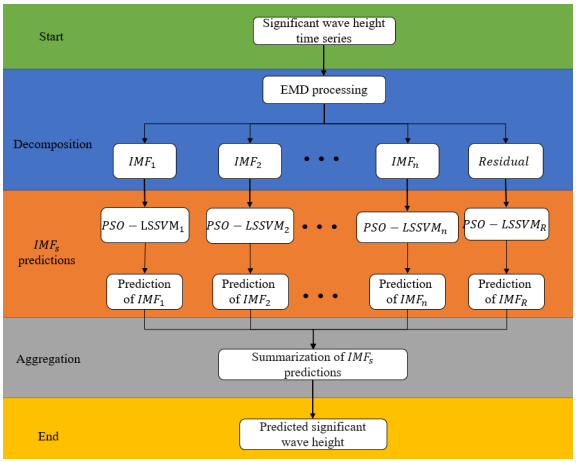


Figure 1. Flowchart of the hybrid EMD-PSO-LSSVM prediction model. The flow chart includes three important steps of our prediction model: EMD data preprocessing, PSO-LSSVM prediction, and finally significant wave height prediction.

2.2 Preprocess data by EMD

Empirical mode decomposition (EMD) is an empirical analysis tool used for processing nonlinear and nonstationary datasets. The main idea of EMD is to decompose the nonlinear and nonstationary time series into a sum of several simple intrinsic mode function (IMF) components and one residue with individual inherent time scale properties. Each IMF represents a kind of natural oscillatory mode and has to satisfy the following two conditions.

(1) the number of extremes and the number of zero-crossings should be equal or differ by one, (2) and the local average should be null, i.e., the mean of the upper envelope defined by the local maxima and the lower envelope defined by the local minima is null.

With a given significant wave height time sequence $x(t)$, EMD processing steps are summarized as follows.

(1) Identify the local extrema.

(2) Generate the upper envelope $u(t)$ and the lower envelope $l(t)$ via a spline interpolation among all the local maxima and the local minima, respectively. Then, the mean envelope is obtained as follows: $m(t) = [l(t) + u(t)]/2$.

(3) Subtract $m(t)$ from the signal $h(t)$ to obtain the IMF candidate, i.e.,$h(t) = x(t) - m(t)$.

(4) Verify whether $h(t)$ satisfies the conditions for IMFs, do step (1) to step (4) until $h(t)$ is an IMF.

(5) Get the *n*th IMF component $imf_n = h(t)$ (after *n* shifting processes) and the corresponding residue $r(t) = x(t) - h(t)$.

(6) Repeat the whole algorithm with $r(t)$ obtained in step (5) until the residue is a monotonic function.

(7) By implementing these algorithms, the decomposition procedure of a signal is expressed as

$$x(t) = \sum_{i=1}^{n} imf_i(t) + r(t)\,. \tag{1}$$

2.3 Least square support vector machine (LSSVM)

Support vector machine (SVM) is a statistical learning theory-based method with a strong capacity to handle nonlinear problems. Its basic idea is to map the nonlinear data into a high dimensional feature space using a nonlinear mapping function, where linear techniques are available. LSSVM is the least squares formulation of a standard SVM. Unlike the inequality constrains introduced in the standard SVM, LSSVM proposed equality constrains in the formulation. This makes the solution being transformed from one of solving a quadratic program to a set of linear equations known as the linear Karush-Kuhn-Tucker (KKT) systems. LSSVM is a nonlinear prediction model based on SVM theory, and it has been widely applied in short-term prediction problems. The LSSVM is retained in this paper as it has good ability for data generalization. It has been shown that the results of a LSSVM model in the prediction problem are also better than other nonlinear models. The basic idea of the method can be described as follows.

Given a training data set of N points $\{(x_i, y_i), i = 1, 2, \ldots, N\}$ with input data $x_i \in R^N$ and output data $y_i \in R$. Define a nonlinear mapping function to map the input data into the high dimensional feature space. In the high dimensional feature space, there theoretically exists a linear function to express the nonlinear relationship between input and output data. Such a linear function, namely the LSSVM function, can be defined as

$$y(x) = \omega^T \phi(x_i) + b, \tag{2}$$

where ω and b are adjustable coefficients. The corresponding optimization problem for LSSVM is formulated as

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Ocean Science
Discussions
Open Access
EGU

$$\begin{cases} Min \;\; J(\omega, e_i) = \frac{1}{2}\|\omega\|^2 + \frac{1}{2}C\sum_{i=1}^{N} {e_i}^2 \\ y(x) = \omega^T \phi(x_i) + b + e_i i \end{cases}, \tag{3}$$

where C denotes the regularization constant and e_i represents the training data error.

The Lagrangian is represented by

$$L(\omega, a_i, b, e_i) = J + \sum_{I=1}^{N} a_i[y_i - \omega^T \phi(x_i) - b - e_i i]. \tag{4}$$

From the Karush-Kuhn-Tucker (KKT) conditions, the following equations must be satisfied

$$\frac{\partial L}{\partial \omega} = 0; \frac{\partial L}{\partial a_i} = 0; \frac{\partial L}{\partial b} = 0; \frac{\partial L}{\partial e_i} = 0. \tag{5}$$

The solution is found by solving the system of linear equations expressed in the following matrix form

$$\begin{bmatrix} 0 & 1_v^T \\ 1 & \Psi + C^{-1}I \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \tag{6}$$

with $y = [y_1, \ldots, y_N]^T, 1_v = [1, \ldots, N]^T, a = [a_1, \ldots, a_N]^T$, I is the identity matrix.$\Psi_{ij} = K(x_i, x_j), i, j = 1, \ldots, N$,which satisfies Mercer's condition.

The LSSVM regression model becomes

$$f(x) = \sum_{i=1}^{n} a_i K(x_i, x_j) + b\,, \tag{7}$$

where a_i are the Lagrange multipliers that can be got by solving the dual problem and $K(x_i, x_j)$ is the kernel function that equals the inner product of $\phi(x_i)$ and $\phi(x_j)$.

The most frequently used kernel functions are the polynomial kernel function, sigmoid kernel function, and radial basis kernel function (RBF). Considering that the RBF kernel is not only easy to implement but also an efficient tool for dealing with nonlinear problems, we selected and retained the RBF function, which is defined by the following equation:

$$K(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right). \tag{8}$$

It is well known that the efficiency of the LSSVM generalization (prediction accuracy) depends on a good collection of meta parameters, parameters C, σ, and parameters of the kernel. When the RBF function is selected, the parameters (C and σ) must be optimized using the PSO-LSSVM system. The regularization parameter C and kernel parameter σ of LSSVM have a significant influence on the classification accuracy. The choices of C and σ govern the model complexity of the prediction.

2.4 LSSVM optimization by PSO

To avoid the under-fitting and over-fitting issues, the LSSVM model's hyper-parameters should be appropriately tuned. This paper uses the particle swarm optimization (PSO) algorithm to find the best value of C and σ in LSSVM. The LSSVM fitting process optimized by particle swarm algorithm is shown in Fig. 2.

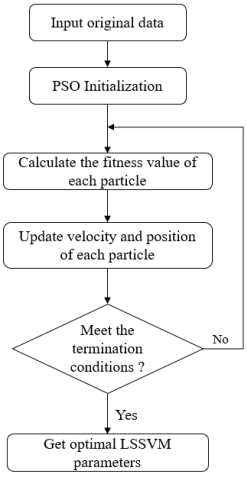


Figure 2. Flowchart of PSO-based parameter selection algorithm. Using the PSO algorithm to find the best value of C and σ in LSSVM.

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Particle swarm optimization uses the velocity-position search model. The iteration formula adjusting the position and speed of the particle is as follows:

$$V_i^{t+1} = wV_i^t + c_1 r_1 (p_{best} - X_i^t) + c_2 r_2 (g_{best} - X_i^t), \quad (9)$$

$$X_i^{t+1} = X_i^t + V_i^{t+1}, \quad (10)$$

where w is the inertial weight; c_1 and c_2 are cognition and social learning factor respectively; $r_1$and r_2 are two random numbers; t denotes the tth iteration; X_i^t is the position of the particle i in d-dimensional space, which denotes the current value of LSSVM parameters C and σ; V_i^t denotes the velocity of a particle i in d-dimensional space, which decides to update the direction and distance of the next generation of C and σ; p_{best} is the best position that every particle can be got during the execution of the PSO method; g_{best} is the best situation that particles have obtained during the implementation of the PSO method.

The following are some parameter descriptions and parameter settings of the particle swarm algorithm.

The iteration t is set to 50, c_1 and c_2 are the cognition and social learning factor respectively, their default values are set as 1, and they can ensure that particles are more affected locally or globally. $r_1$and r_2 are two random numbers in the range $[0, 1]$. The use of the inertial weight controls the previous history of velocity on the current one. w is the weight factor. A considerable inertia weight facilitates global exploration, while a small one tends to facilitate local exploration. A suitable value of the inertia weight usually provides a balance between the global and regional exploration abilities. We used a linearly decreasing inertia weight, which starts at 0.9 and ends at 0.4, the performance of PSO can be significantly improved. The inertial weight can be expressed as follows:

$$w = w_{max} - \frac{w_{max} - w_{min}}{t_{max}} \times t, \quad (11)$$

where t_{max} is the maximum iteration counter. w_{max} and w_{min} are the initial and terminal weights, respectively. New fitness values of the particles are calculated after the velocity and position updates if required, p_{best} and g_{best} are also updated, and the same procedure is performed continuously until the stop criteria are satisfied. Usually, each particle's velocity is restricted to a maximum value within the interval $[-0.01, 100]$, which is defined according to the bounds on decision variables.

3 Descriptions of the wave data and prediction accuracy measures

3.1 Raw data

Two North Atlantic Ocean areas have been selected to predicted the significant wave height. The significant wave height and meteorological series were downloaded from National Data Buoy Center (NDBC) (https://www.ndbc. noaa.gov). Two stations were utilized in this study (Fig. 3), point A is station 41025 at 35°1'30" N 75°21'47" W, in the offshore, while point B is station 41048 at 31°49'53" N 69°34'23" W, in the deep-sea zone. These stations were selected as they have an unimpaired and long series of recorded significant wave height and metrological data.

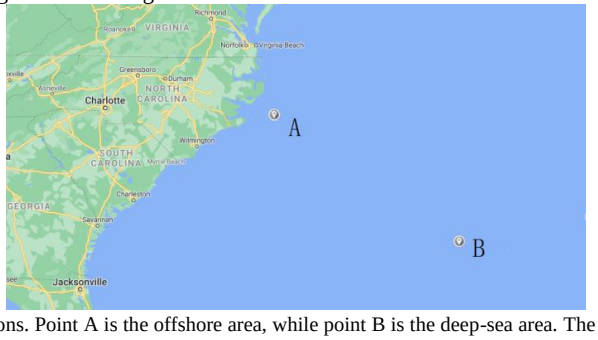


Figure 3. Location map of the stations. Point A is the offshore area, while point B is the deep-sea area. The map download from © Google Maps, we added two marks for our study locations.

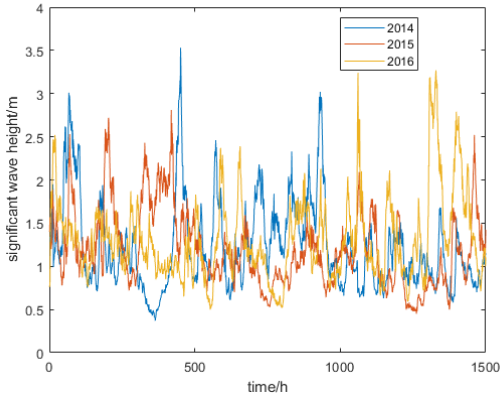
There are three sections of used data. The two sites in 2014, 2015, and 2016 are partially significant wave height data, with 1500 sample points taken out each year. Data from 2014 and 2015 were used as training data and data from 2016 were used as testing data. Fig. 4 shows the significant wave height (SWH) records of both points. Table 1 shows the minimum, maximum, and average values of different training parameters and testing data sets.

Table 1. Specific information on the two stations.

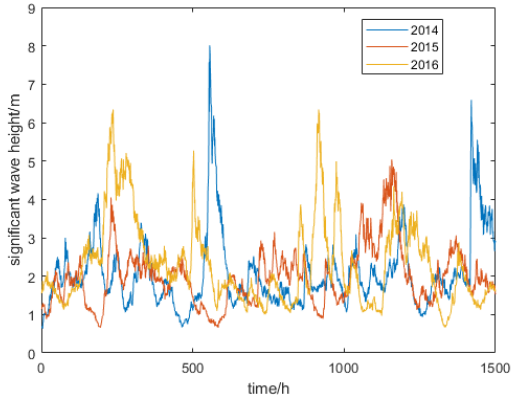
Station ID	Water depth(m)	Year	SWH average(m)	SWH range(m)
41025	59.4	2014	1.2287	[0.37,3.53]
		2015	1.1938	[0.45,2.81]
		2016	1.3141	[0.50,3.27]
41048	5309	2014	2.0715	[0.63,8.01]
		2015	1.9668	[0.66,5.04]
		2016	2.2987	[0.67,6.34]

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Ocean Science
Discussions
Open Access
EGU



a)



b)

Figure 4. Significant wave height of two points. a) 41025, b) 41048. Data from 2014 and 2015 were used as training data and data from 2016 were used as testing data.

As can be seen from Table 1 and Fig. 4, the average significant wave height at station 41025, located near the coast, is around 1.2 m, with the maximum significant wave height around 3 m. The sea state is relatively stable. The average significant wave height at station 41048, located in the deep-sea area, is about 2 m, and the maximum significant wave height is about 6.5 m. The sea conditions are relatively rough. Therefore, it is difficult to predict the significant wave height in the deep-sea area.

The significant wave height data for 2014 and 2015 were used as input variables for model development. The relevance of each feature with significant wave height needs to be determined before choosing the input features. The correlation coefficient $r_{x,y}$ can be calculated as

$$r_{x,y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad (12)$$

where $r_{x,y}$ represents the correlation coefficient between data set x and y, i is a positive integer. $|r| \geq 0.8$ indicates that there is a high correlation between the two features. The correlation coefficient of the input features with the output feature is shown in Table 2. *H-i* in the table represents the significant wave height data from ith hours ago, *H-2* represents the significant wave height data from two hours ago, as an example. From the table, it can be seen that the correlation coefficient is lower than 0.8 at *H-6*, so the data from five hours ago are used as input in this paper.

Table 2. Correlation coefficient of the input features with output feature.

	H-1	H-2	H-3	H-4	H-5	H-6
$r_{x,y}$	0.9707	0.9435	0.9093	0.8710	0.8039	0.7680

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Ocean Science
Discussions
Open Access
EGU

3.2 Models evaluations

To evaluate the performance of the models, statistical and standardized metrics were used. The mathematical formulations of these assessment metrics are given as follows.

1) Root mean square error (RMSE) is expressed as follow:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}, \tag{13}$$

2) Mean absolute error (MAE) is expressed as follow:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|, \tag{14}$$

3) Mean square error (MSE) is expressed as follow:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2, \tag{15}$$

4) Coefficient of determination (R^2) is expressed as follow:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{16}$$

where x and y are the observed and the predicted values, respectively; $\bar{x}$ is the mean values of the observed values; n is the number of observations. The low values of RMSE, MAE, and MSE reveal the acceptable accuracy of the models. The R^2 ranges between 0 and 1, where 1 indicates a perfect positive linear relationship between the observed and the predicted values, and 0 shows no relationship.

4 Results and discussion

4.1 Single models

We first consider the single model to predict the significant wave height. The single model used here is LSSVM, ELM, and ANN, and the significant wave height is predicted for 1 hour and 3 hours. The specific parameters of various model networks are shown in Table 3. IN is the number of input layer units, H is the number of hidden layer units, O is the number of output layer units, σ is the confidence, C is the penalty coefficient.

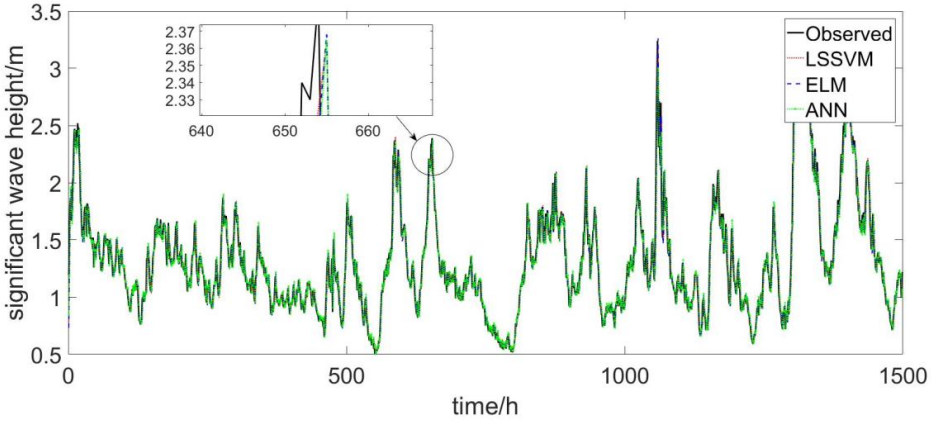
Table 3. Parameter of three single models.

Model	Initial settings
LSSVM	IN=5, O=1, $\sigma = 10, C$=100, kernel function = Radial Basis Function (RBF)
ELM	IN=5, H=10, O=1, activation function = Sigmoid
ANN	IN=5, H=10, O=1, training algorithm = Levenberg-Marquardt

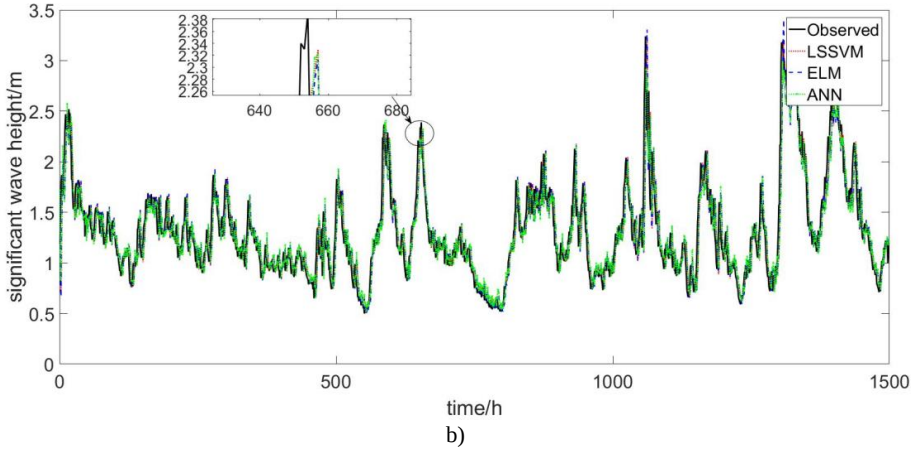
Fig. 5 and Fig. 6 show the predictions of the significant wave height of 41025 and 41048 stations by three single models. Table 4 shows the numerical analysis of specific evaluation indicators. It can be seen from Table 3 that for the wave height prediction of 41025 stations near the coast, R^2 can be kept above 0.8 when the 3-hour prediction is made. For the 41048 station in the deep-sea area, R^2 can be maintained above 0.9 during the 3-hour prediction, and there is a high correlation between the predicted significant wave height and the observed significant wave height. In general, the three algorithms have achieved satisfactory results in predicting the significant wave height, but LSSVM has higher prediction accuracy than the other two models. This clearly shows that compared with other models, the proposed LSSVM model can be considered as the best wave height predictor.

It can be seen from Fig. 5 and Fig. 6 that the observed wave height and the predicted wave height are slightly misaligned on the time scale axis. It can be seen from the enlarged view in the figure that a one-time step significantly shifts the predicted wave heights of the three single models. These wave forecasting models exhibit lag in the prediction timing, making the univariate time series forecasting a futile attempt. As the leading time increases, these lags become larger. The lag is a type of prediction error that can also be found in other work on wave forecasting using single models. The lag mainly results from the nonstationarity hidden in the measured wave time series. Modeling a nonlinear and nonstationary data set by applying a single nonlinear model is very difficult because there are too many possible patterns hidden in the data. A single model may not be general enough to capture all the essential features. Even if the nonlinear ANN is used to forecast the nonlinear and nonstationary wave heights, the lags remain. A single prediction model cannot capture all the components with different scales simultaneously. Therefore, the "lags" occur in the forecasting results. This would give a direct and physical explanation. These apparent lag phenomena affect the accuracy of prediction, so the following work content is to eliminate this lag phenomenon and improve the prediction accuracy.

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

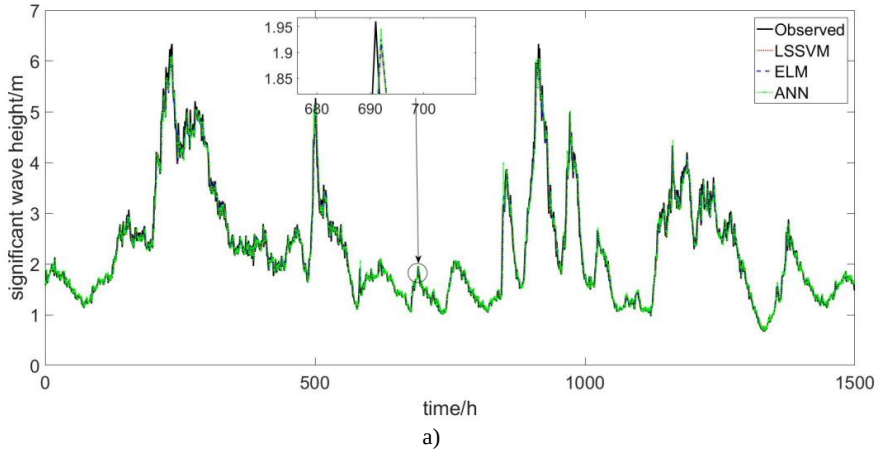


a)



b)

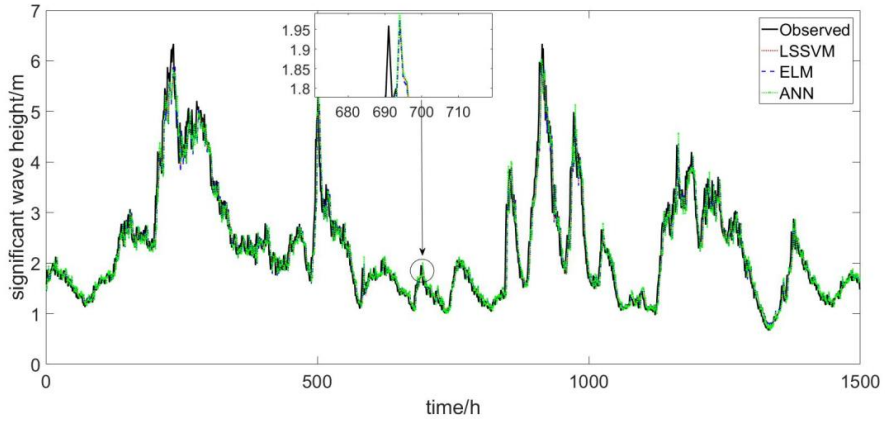
Figure 5. Comparison between the observed and predicted significant wave height at station 41025. a) one hour, b) three hours. For station 41025, in the near-coastal area, there is a time lag in the prediction using a single model. Comparing a) and b), it can be seen that as the leading time increases, the prediction timing lag problem becomes greater.Figure.5



a)

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Ocean Science
Discussions
Open Access
EGU



b)

Figure 6. Comparison between the observed and predicted significant wave height at station 41048. a) one hour, b) three hours. For station 41048, in the deep-sea area, there is a time lag in the prediction using a single model. Comparing a) and b), it can be seen that as the leading time increases, the prediction timing lag problem becomes greater.

Table 3 Analysis of the prediction results of three single models

Station	Model	Leading time	RMSE	MAE	MSE	R^2
41025	LSSVM	1	0.115	0.082	0.013	0.942
		3	0.200	0.141	0.040	0.826
	ELM	1	0.115	0.082	0.013	0.942
		3	0.201	0.141	0.040	0.826
	ANN	1	0.116	0.082	0.013	0.942
		3	0.201	0.141	0.041	0.825
41048	LSSVM	1	0.183	0.126	0.033	0.972
		3	0.276	0.184	0.076	0.936
	ELM	1	0.184	0.127	0.034	0.972
		3	0.279	0.185	0.079	0.936
	ANN	1	0.188	0.128	0.035	0.971
		3	0.278	0.184	0.077	0.935

4.2 Hybrid models

The time series of ocean waves is a complicated nonlinear and nonstationary signal that consists of different oscillation scales. The time series prediction of non-stationary data by using single models only will lead to the homogenization of the original input data's various characteristics, which could affect the prediction accuracy and cause the lag phenomenon. Accordingly, the non-stationarity of the time series of significant wave height and input variables should be reduced. The combination of an EMD model with an LSSVM model provides an effective way to improve the wave prediction. The EMD was adopted to decompose a significant wave height series that consisted of one residual series and several IMFs. Then, the residual series and IMFs were modeled by the LSSVM model, and finally the summation of the prediction output of subseries significant wave height. In addition, the PSO was employed to optimize the LSSVM parameters to increase the prediction accuracy.

In the first step, the wave height time series is decomposed into a couple of meaningful and straightforward IMFs and a residual by EMD (Fig. 7).

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Ocean Science
Discussions
Open Access
EGU

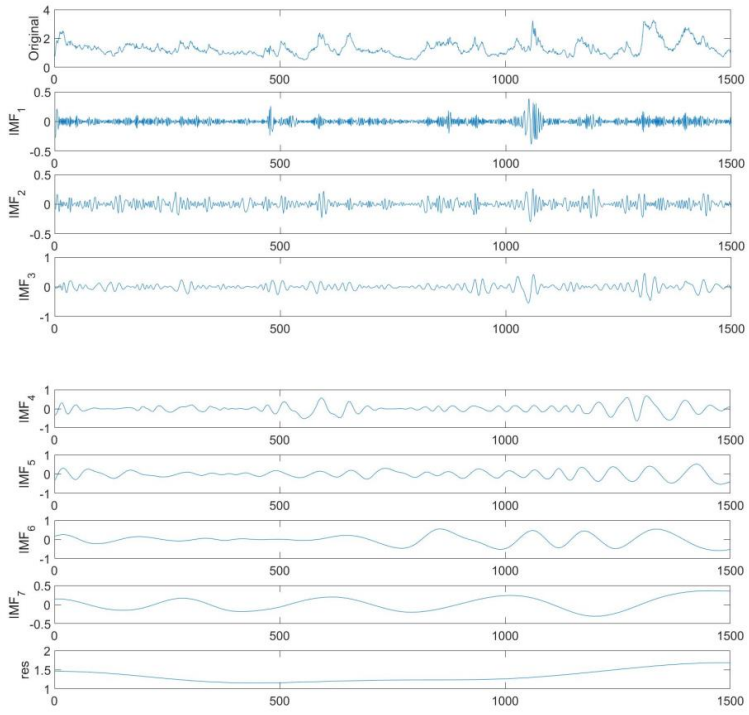


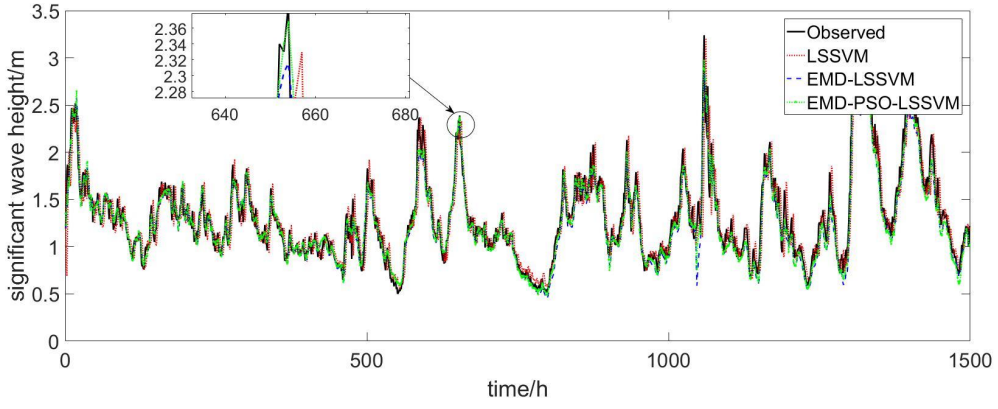
Figure 7. Decomposition results of significant wave height time series data using the EMD. The EMD decomposition decomposes the nonlinear significant wave height into 7 IMFs and 1 res, which reduces the difficulty of prediction.

Significant wave data sets are decomposed into IMFs and residuals when implementing the EMD-based prediction models. Fig. 7 displays the decomposition results of wave height time series measured at Station 41025, the EMD decomposition decomposes the nonlinear significant wave height into 7 IMFs and 1 res, where it is seen that several simple components can represent the complex wave height time series. This would have enabled the single model to extract features during the modeling of significant wave height effectively. Next, an EMD-based hybrid model will be used to predict the significant wave height.

As can be seen in Fig. 8, the single model LSSVM shows a prediction timing lag (red dotted line).The other two models have overcome the lag by using the EMD technique, prediction results for the nonlinear and nonstationary waves were improved mainly by combining the EMD technique with the single model.

It appears that the use of the EMD decomposition hybrid model solves the lag phenomenon and improves the prediction accuracy. As shown in Table 4, the comparison between the EMD-LSSVM and LSSVM shows that the EMD-LSSVM model is superior to the single model LSSVM in various numerical values. Especially for long-term predictions, taking a prediction leading time of 6 hours as an example, the R^2 of the EMD-LSSVM model for stations 41025 and 41048 are 0.888 and 0.959, respectively, and the R^2 of the LSSVM model for stations, 41025 and 41048 are 0.645 and 0.858, respectively. The correlation based on the EMD-LSSVM model prediction is significantly higher, while the RMSE, MAE, and MSE are reduced considerably.

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021



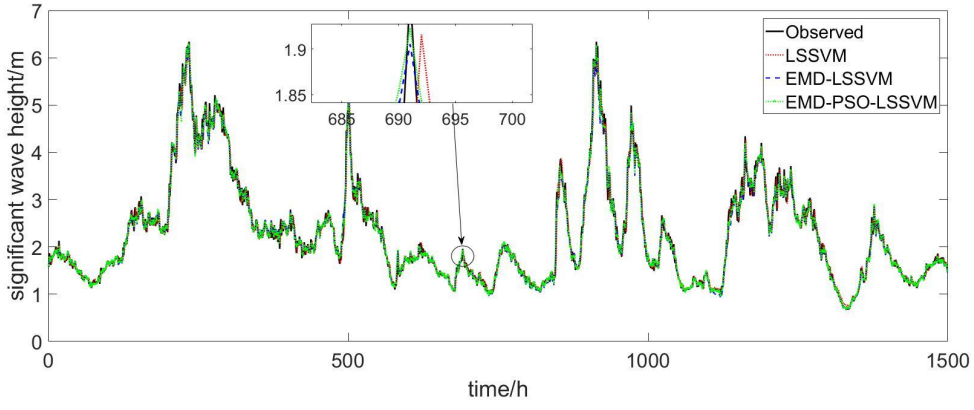


Figure 8. Comparison between LSSVM, EMD-LSSVM and EMD-PSO-LSSVM. a) 41025, b) 41048. By adding the EMD method to preprocess the data, the new hybrid model removed the time lag of a single model. At the same time, the prediction accuracy of the peak of the significant wave height has also been improved.

It can be seen from Fig. 8 that the preprocessing method of EMD decomposition has solved the lag phenomenon. However, the significant wave height prediction effect using the LSSVM model is still not very satisfactory. For example, there are errors in predicting the peaks and troughs of the significant wave height. The next step is to optimize LSSVM parameters to improve the prediction accuracy of the model.

Changing the parameter values of a prediction system can have a significant impact on its performance. Therefore, we should find the optimum parameter values for the prediction system , human experts have performed this task, who typically use a priori knowledge to specify the parameter values. However, this approach can be subject to human bias. PSO has emerged as a practical tool for high-quality parameter selection in prediction systems.

PSO is used to optimize the LSSVM parameters. The methodological steps can be found in the description of the method in section 2. Fig. 8 shows that the EMD-PSO-LSSVM model can predict the significant wave height peaks and troughs very well, significantly improving the prediction accuracy.

Table 4 presents the results obtained of the significant wave height prediction by the EMD-PSO-LSSVM method. As can be seen, the prediction of significant wave height is accurate with the proposed method, and the effect of using PSO to optimize the LSSVM parameters can be seen with an improvement of the prediction accuracy from R^2=0.972, 0.945 and 0.888 (EMD-LSSVM) to R^2 =0.972, 0.958 and 0.902 (EMD-PSO-LSSVM) at station 41025, and an improvement of the prediction accuracy from R^2 =0.991, 0.982 and 0.959 (EMD-LSSVM) to R^2 =0.993, 0.987 and 0.965 (EMD-PSO-LSSVM) at station 41048. Correspondingly, the RMSE, MAE, and MSE predicted by EMD-PSO-LSSVM at the two sites are also the lowest. With increasing prediction leading time, the model performed well using the improved EMD-PSO-LSSVM model, especially at station 41048 in the deep-sea region.

It can be seen from Table 4 that the prediction effect at station 41048 is better than that at station 41025. One of the reasons for

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Ocean Science
Discussions
Open Access
EGU

this is that 41025 is a station near the coast. The significant wave height is relatively stable, and there are few big winds. In extreme cases, the training data is less. It can be seen from Fig. 9 and Fig. 10 that the significant wave height is relatively stable. When it is higher, the fitted data points are more scattered. 41048 is in the deep-sea area, the significant wave height range is relatively large, and the training data is also relatively large. It can be seen from Fig. 11 and Fig. 12 that when the significant wave height is high, the fitting effect of the data points is still better.

Table 4. Performance results for station 41025 and station 41048.

Station	Algorithm	Leading time	RMSE	MAE	MSE	R^2
41025	EMD-PSO-LSSVM	1	0.089	0.062	0.008	0.972
		3	0.105	0.079	0.011	0.958
		6	0.155	0.112	0.024	0.902
	EMD-LSSVM	1	0.097	0.071	0.009	0.972
		3	0.125	0.092	0.016	0.945
		6	0.169	0.123	0.029	0.888
	LSSVM	1	0.115	0.082	0.013	0.942
		3	0.200	0.141	0.040	0.826
		6	0.287	0.202	0.082	0.645
41048	EMD-PSO-LSSVM	1	0.089	0.063	0.008	0.993
		3	0.127	0.091	0.016	0.987
		6	0.205	0.140	0.042	0.965
	EMD-LSSVM	1	0.105	0.074	0.011	0.991
		3	0.150	0.104	0.022	0.982
		6	0.224	0.154	0.050	0.959
	LSSVM	1	0.183	0.126	0.034	0.972
		3	0.278	0.184	0.076	0.936
		6	0.416	0.277	0.173	0.858

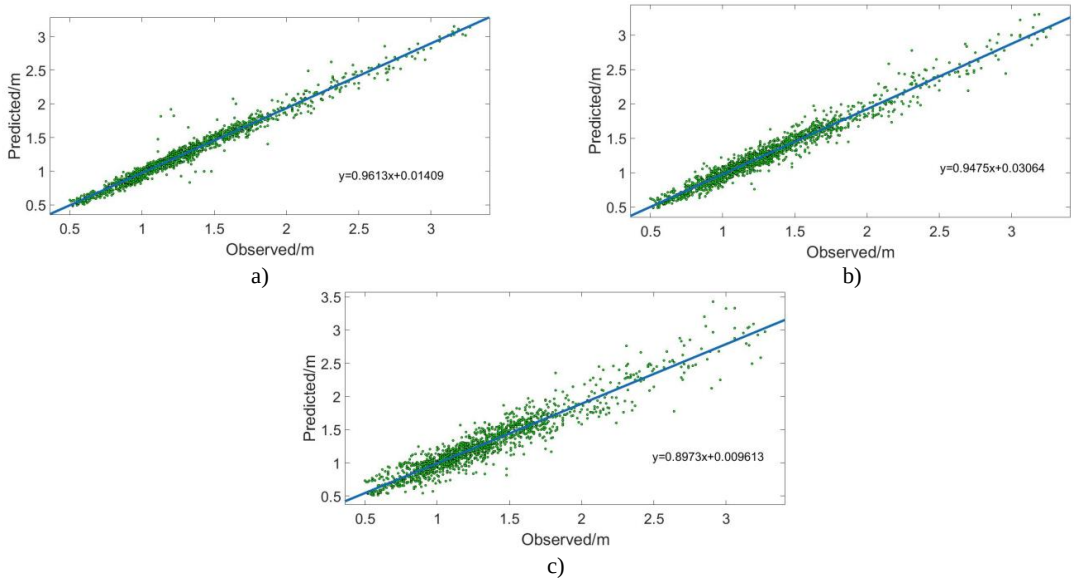


Figure 9. Scatter diagram of the observed and the predicted using the EMD-PSO-LSSVM at Station 41025. a) one hour, b) three hours, c) six hours. The predicted value and the observed value are linearly fitted. The higher the slope of the fitted line, the better the prediction effect. The fitting slopes of EMD-PSO-LSSVM to 41025 stations at 1, 3, and 6 hours are 0.9613, 0.9475, and 0.8973.

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Ocean Science
Discussions
Open Access
EGU

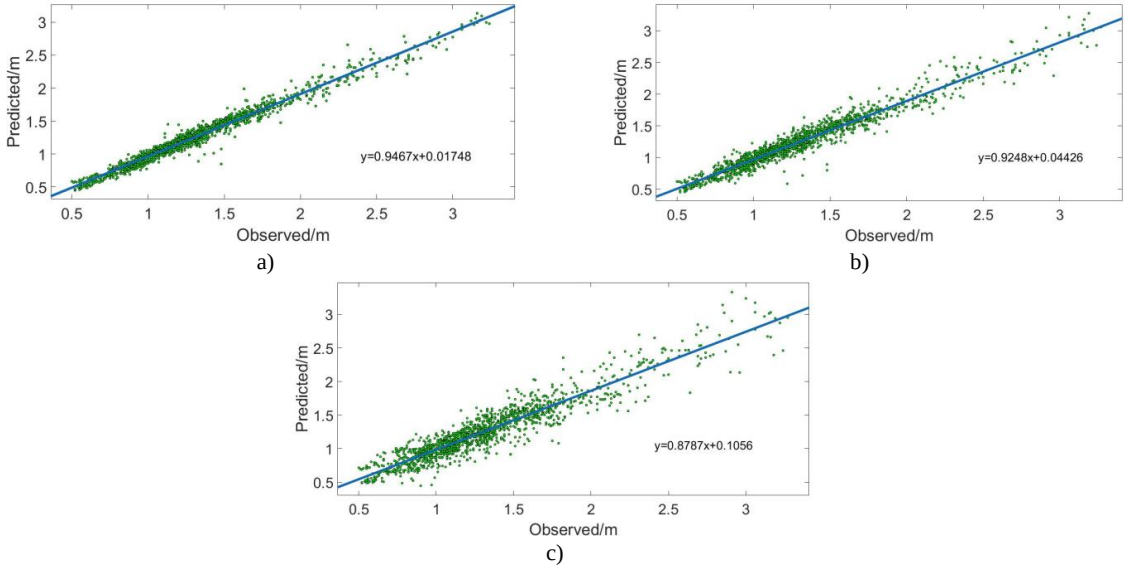


Figure 10. Scatter diagram of the observed and the predicted using the EMD- LSSVM at Station 41025. a) one hour, b) three hours, c) six hours. The fitting slopes of EMD-PSO-LSSVM to 41025 stations at 1, 3, and 6 hours are 0.9467, 0.9248, and 0.8787. Compared with the hybrid model using PSO for parameter optimization, the accuracy of the model is not as good as the improved model, so the effectiveness of the PSO method is proved.

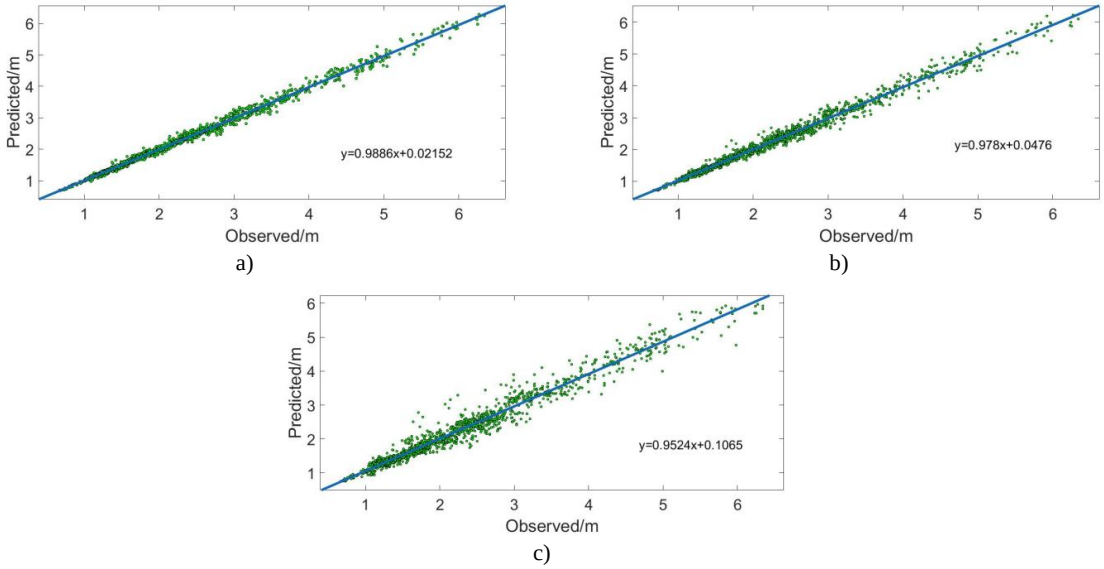


Figure 11. Scatter diagram of the observed and the predicted using the EMD-PSO-LSSVM at Station 41048. a) one hour, b) three hours, c) six hours. The fitting slopes of EMD-PSO-LSSVM to 41048 stations at 1, 3, and 6 hours are 0.9886, 0.978, and 0.9524. Compared with 41025 in the offshore, the combined model is more suitable for deep-see area, and the fitted slopes are all above 0.95.

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

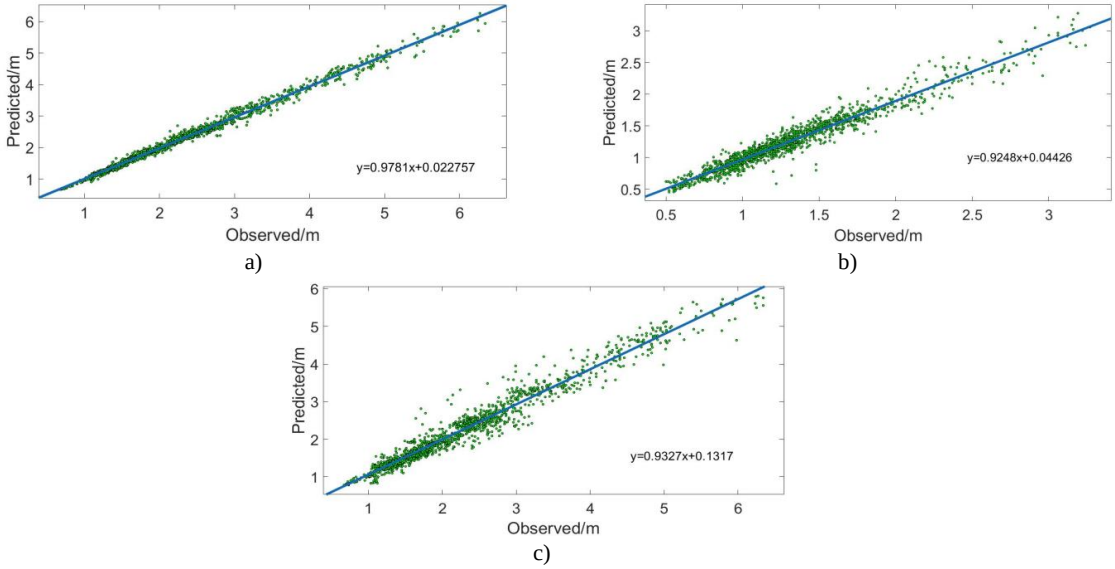


Figure 12. Scatter diagram of the observed and predicted using the EMD- LSSVM at Station 41048. a) one hour, b) three hours, c) six hours. The fitting slopes of EMD-PSO-LSSVM to 41048 stations at 1, 3, and 6 hours are 0.9781 0.9428, and 0.9327. As with the results for station 41025, the hybrid model using PSO has higher accuracy.

Fig. 9 and Fig. 11 show a comparison between the observed and predicted values by EMD-PSO-LSSVM at station 41025 and 41048, respectively. Fig. 10 and Fig. 12 show a comparison between the observed and predicted values by EMD-LSSVM at station 41025 and 41048, respectively. Referring to the scatter plots, by looking at the distribution of the scatter plot and the slope of the fitted line, we can clearly see that the relationship between the predicted and observed values. The denser the scatter plot distribution and the closer the slope of the fitted line is to 1, and the better is the prediction result. From those figures, it appears that the estimations of EMD-LSSVM time series are more scattered and farther than those EMD-PSO-LSSVM model. As the leading time increases, EMD-LSSVM performance decreases drastically, but EMD-PSO-LSSVM performance decreases gradually, as shown in figures. For example, the best-fit line slope for the scatter of wave prediction of six leading hours at Station 41025 and 41048 are 0.8973 and 0.9524, respectively. Comparatively, the EMD-PSO-LSSVM model performs better than the EMD-LSSVM model. The coefficient of determination for the wave prediction at all stations using the EMD-PSO-LSSVM model is more than 0.902 (see Table 4), while the best-fit line slopes for the scatters are better than 0.8973 (Fig. 9 and Fig. 11). Correspondingly, the RMSE, MAE, and MSE predicted by EMD-PSO-LSSVM at the two sites are also the lowest.

5 Conclusions

This paper introduces a new prediction method, using EMD-PSO-LSSVM for nonlinear and nonstationary significant wave height prediction. It has high adaptability and accuracy for dealing with any random time series wave prediction problem.

We have carried out some actual forecasting operations on the significant wave heights in the offshore and deep-sea areas of the North Atlantic Ocean, using single models and hybrid models to make the prediction. Several statistical indices were utilized for evaluating the accuracy of the predictions of the proposed models. From the obtained results, due to the nonlinearity and nonstationarity of the significant wave height, the traditional single models have the phenomenon of lagging prediction; and as the leading time increases, the lag in the prediction becomes more and more serious. This lagging phenomenon reduces the prediction accuracy, and of course, it will impact the actual engineering applications. Therefore, the EMD method is added to preprocess the significant wave height based on the time series, and the EMD-LSSVM hybrid model with preprocessing can well solve the problem of prediction lag. However, the predicted results are not very satisfactory. For example, the prediction of the peaks and troughs of the significant wave height is not accurate, which reduces the prediction accuracy. Therefore, the PSO algorithm is added to the original EMD-LSSVM hybrid method, and the critical parameters of LSSVM are optimized through the PSO algorithm. In this way, a new hybrid model EMD-PSO-LSSVM is proposed. Significant wave height data from two NDBC buoys with various geographical and statistical properties were used in the comparison studies. Various data from Table 4 show that the proposed EMD-PSO-LSSVM predictor of significant wave height in different locations show consistent conclusions. The proposed new hybrid model can solve the problem of prediction lag and significantly improve the prediction accuracy.

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Ocean Science Discussions

Data availability: Data used in this paper were downloaded from the National Data Buoy Center (NDBC). The datasets used were: https://www.ndbc.noaa.gov/station_history.php?station=41025 https://www.ndbc.noaa.gov/station_history.php?station=41048

Author Contributions: Conceptualization, G.T. and H.D.; methodology, H.D.; writing—original draft preparation, H.D.; writing—review and editing, C.C. and Y.W.; visualization, X.H.; supervision, G.T.; project administration, G.T.; funding acquisition, S.M. All authors have read and agreed to the published version of the manuscript.

Acknowledgements: This work was supported in part by the Medical Science and Technology Research Foundation of Guangdong under Grant A2020334, in part by the Youth Creative Talent Project (Natural Science) of Guangdong under Grant 2019KQNCX018, in part by the Young Talent Training Project of Guangzhou University of Chinese Medicine under Grant QNYC20190110.
Conflicts of Interest: The authors declare no conflict of interest.

References

Agrawal, J. D., and Deo, M. C.: On-line wave prediction, Marine Structures, 15, 57-74, https://doi.org/10.1016/S0951-8339(01)00014-4, 2002.

Akbarifard, S., and Radmanesh, F.: Predicting sea wave height using Symbiotic Organisms Search (SOS) algorithm, Ocean Engineering, 167, 348-356, https://doi.org/10.1016/j.oceaneng.2018.04.092, 2018.

Browne, M., Castelle, B., Strauss, D., Tomlinson, R., Blumenstein, M., and Lane, C.: Near-shore swell estimation from a global wind-wave model: Spectral process, linear, and artificial neural network models, Coastal Engineering, 54, 445-460, https://doi.org/10.1016/j.coastaleng.2006.11.007, 2007.

Casas-Prat, M., Wang, X. L., and Sierra, J. P.: A physical-based statistical method for modeling ocean wave heights, Ocean Modelling, 73, 59-75, https://doi.org/10.1016/j.ocemod.2013.10.008, 2014.

Chen, J., Heincke, B., Jegen, M., and Moorkamp, M. J. G. J. I.: Using empirical mode decomposition to process marine magnetotelluric data, 190, 293-309, 2012.

Cornejo-Bueno, L., Garrido-Merchán, E. C., Hernández-Lobato, D., and Salcedo-Sanz, S.: Bayesian optimization of a hybrid system for robust ocean wave features prediction, Neurocomputing, 275, 818-828, https://doi.org/10.1016/j.neucom.2017.09.025, 2018.

Deka, P. C., and Prahlada, R.: Discrete wavelet neural network approach in significant wave height forecasting for multistep lead time, Ocean Engineering, 43, 32-42, https://doi.org/10.1016/j.oceaneng.2012.01.017, 2012.

Deo, M. C., and Sridhar Naidu, C.: Real time wave forecasting using neural networks, Ocean Engineering, 26, 191-203, https://doi.org/10.1016/S0029-8018(97)10025-7, 1998.

Deo, M. C., Jha, A., Chaphekar, A. S., and Ravikant, K.: Neural networks for wave forecasting, Ocean Engineering, 28, 889-898, https://doi.org/10.1016/S0029-8018(00)00027-5, 2001.

Dixit, P., Londhe, S., and Dandawate, Y.: Removing prediction lag in wave height forecasting using Neuro - Wavelet modeling technique, Ocean Engineering, 93, 74-83, https://doi.org/10.1016/j.oceaneng.2014.10.009, 2015.

Duan, W.-y., Huang, L.-m., Han, Y., and Huang, D.-t.: A hybrid EMD-AR model for nonlinear and non-stationary wave forecasting, Journal of Zhejiang University-SCIENCE A, 17, 115-129, 10.1631/jzus.A1500164, 2016a.

Duan, W. Y., Han, Y., Huang, L. M., Zhao, B. B., and Wang, M. H.: A hybrid EMD-SVR model for the short-term prediction of significant wave height, Ocean Engineering, 124, 54-73, https://doi.org/10.1016/j.oceaneng.2016.05.049, 2016b.

Etemad-Shahidi, A., and Mahjoobi, J.: Comparison between M5′ model tree and neural networks for prediction of significant wave height in Lake Superior, Ocean Engineering, 36, 1175-1181, https://doi.org/10.1016/j.oceaneng.2009.08.008, 2009.

Fan, S., Xiao, N., and Dong, S.: A novel model to predict significant wave height based on long short-term memory network, Ocean Engineering, 205, 107298, https://doi.org/10.1016/j.oceaneng.2020.107298, 2020.

https://doi.org/10.5194/os-2021-2
Preprint. Discussion started: 28 January 2021

Flandrin, P., Rilling, G., and Goncalves, P. J. I. s. p. l.: Empirical mode decomposition as a filter bank, 11, 112-114, 2004.

Guedes Soares, C., and Cunha, C.: Bivariate autoregressive models for the time series of significant wave height and mean period, Coastal Engineering, 40, 297-311, https://doi.org/10.1016/S0378-3839(00)00015-6, 2000.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N.-C., Tung, C. C., Liu, H. H. J. P. o. t. R. S. o. L. S. A. m., physical, and sciences, e.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, 454, 903-995, 1998.

Huang, N. E., and Wu, Z. J. R. o. g.: A review on Hilbert-Huang transform: Method and its applications to geophysical studies, 46, 2008.

Hwang, P. A. J. J. o. G. R. O.: Duration-and fetch-limited growth functions of wind-generated waves parameterized with three different scaling wind velocities, 111, 2006.

Janssen, P. A. E. M.: Progress in ocean wave forecasting, Journal of Computational Physics, 227, 3572-3594, https://doi.org/10.1016/j.jcp.2007.04.029, 2008.

Kaloop, M. R., Kumar, D., Zarzoura, F., Roy, B., and Hu, J. W.: A wavelet - Particle swarm optimization - Extreme learning machine hybrid modeling for significant wave height prediction, Ocean Engineering, 213, 107777, https://doi.org/10.1016/j.oceaneng.2020.107777, 2020.

Mahjoobi, J., and Adeli Mosabbeb, E.: Prediction of significant wave height using regressive support vector machines, Ocean Engineering, 36, 339-347, https://doi.org/10.1016/j.oceaneng.2009.01.001, 2009.

Makarynskyy, O.: Improving wave predictions with artificial neural networks, Ocean Engineering, 31, 709-724, https://doi.org/10.1016/j.oceaneng.2003.05.003, 2004.

Mandal, S., and Prabaharan, N.: Ocean wave forecasting using recurrent neural networks, Ocean Engineering, 33, 1401-1410, https://doi.org/10.1016/j.oceaneng.2005.08.007, 2006.

Prasad, R., Deo, R. C., Li, Y., and Maraseni, T.: Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm, Atmospheric Research, 197, 42-63, https://doi.org/10.1016/j.atmosres.2017.06.014, 2017.

Ra, W., and Whang, I. J. E. L.: Real-time long-term prediction of ship motion for fire control applications, 42, 1, 2006.

Rhif, M., Ben Abbes, A., Farah, I. R., Martínez, B., and Sang, Y. J. A. S.: Wavelet transform application for/in non-stationary time-series analysis: a review, 9, 1345, 2019.

Richter, M., Schaut, S., Walser, D., Schneider, K., and Sawodny, O.: Experimental validation of an active heave compensation system: Estimation, prediction and control, Control Engineering Practice, 66, 1-12, https://doi.org/10.1016/j.conengprac.2017.06.005, 2017.

Soares, C. G., Ferreira, A. M., and Cunha, C.: Linear models of the time series of significant wave height on the Southwest Coast of Portugal, Coastal Engineering, 29, 149-167, https://doi.org/10.1016/S0378-3839(96)00022-1, 1996.

Tsai, C.-P., Lin, C., and Shen, J.-N.: Neural network for wave forecasting among multi-stations, Ocean Engineering, 29, 1683-1695, https://doi.org/10.1016/S0029-8018(01)00112-3, 2002.

Valamanesh, V., Myers, A. T., Arwade, S. R., Hajjar, J. F., Hines, E., and Pang, W.: Wind-wave prediction equations for probabilistic offshore hurricane hazard analysis, Natural Hazards, 83, 541-562, 10.1007/s11069-016-2331-z, 2016.

Yoon, H., Jun, S.-C., Hyun, Y., Bae, G.-O., and Lee, K.-K.: A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer, Journal of Hydrology, 396, 128-138, https://doi.org/10.1016/j.jhydrol.2010.11.002, 2011.