Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

A hybrid EMD-AR model for nonlinear and non-stationary wave forecasting*

Wen-yang DUAN, Li-min HUANG†‡, Yang HAN, De-tai HUANG

(*Department of Shipbuilding Engineering, Harbin Engineering University, Harbin 150001, China*)

†E-mail: huanglimin@hrbeu.edu.cn

Received June 1, 2015; Revision accepted Nov. 23, 2015; Crosschecked Jan. 16, 2016

Abstract: Accurate wave forecasting with a couple of hours of warning time offers improvements in safety for maritime operation-related activities. Autoregressive (AR) model is an efficient and highly adaptive approach for wave forecasting. However, it is based on linear and stationary theory and hence has limitations in forecasting nonlinear and non-stationary waves. Inspired by the capability of empirical mode decomposition (EMD) technique in handling nonlinear and non-stationary signals, this paper describes the development of a hybrid EMD-AR model for nonlinear and non-stationary wave forecasting. The EMD-AR model was developed by coupling an AR model with the EMD technique. Nonlinearity and non-stationarity were overcome by decomposing the wave time series into several simple components for which the AR model is suitable. The EMD-AR model was implemented using measured significant wave height data from the National Data Buoy Center, USA. Prediction results from various locations consistently show that the hybrid EMD-AR model is superior to the AR model. This demonstrates that the EMD technique is effective in processing nonlinear and non-stationary waves.

Key words: Wave forecast, Nonlinear and non-stationary, Autoregressive (AR) model, Empirical mode decomposition (EMD), EMD-AR model

http://dx.doi.org/10.1631/jzus.A1500164 **CLC number:** U66

1 Introduction

Short-term wave prediction at a location is essential in the design and implementation of maritime operations. Reliable prediction allows improvements in safety for conducting operation-related activities, helps offshore structures to avoid the dangers of harsh conditions (Jain and Deo, 2007), and improves the efficiency of wave energy converters (Li *et al.*, 2012). The operational wave forecast has been widely explored for its application value in engineering over recent decades. A large number of wave forecast models have been developed. According to the theoretical differences among the various methods, wave forecast models may be classified into four types of approaches: energy balance equation (EBE)-based models, classical time series models, intelligent-technique-based nonlinear models, and hybrid models.

Conventional numerical models for wave forecasts were based on EBEs. EBE-based numerical models are usually used to forecast waves over a large spatial and temporal domain (The Wamdi Group, 1988; Komen *et al.*, 1994; Janssen, 2008; Sandhya *et al.*, 2014; Tolman, 2014). However, their predictions depend on how precisely the phenomena are expressed in formulations (Mandal and Prabaharan, 2010). In addition, their implementation remains difficult because of the high computational

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 51079032, 51490671, and 11572093) and the International Science and Cooperation Project Sponsored by the National Ministry of Science and Technology of China (No. 2012DFA70420)

ORCID: Wen-yang DUAN, http://orcid.org/0000-0002-7811-4986; Li-min HUANG, http://orcid.org/0000-0002-7944-2754

cost and the unavailability of forcing functions (Londhe and Panchang, 2006).

Classical time series models provide another possible way for achieving wave forecasting. Their simple model methodology without wind-to-wave conversion does not require exogenous data and massive computing memory and time (Jain and Deo, 2007). However, classical time series models are not suitable for nonlinear and non-stationary wave prediction because of their linear and stationary assumptions. Therefore, several improved models, such as the bilinear model (Hannan, 1982), the threshold autoregressive model (Tong and Lim, 1980), and the autoregressive conditional heteroskedasticity model (Engle, 1982), have been developed. These nonlinear models are limited by the hypothesized explicit relationships for the available data series (Zhang *et al.*, 1998).

To resolve the nonlinearity of ocean waves, intelligent-technique-based nonlinear models such as artificial neural networks (ANNs) and genetic programming (GP) models have been extensively studied. Real-life applications of these soft computing techniques can be found in different fields (Chau, 2007; Wu and Chau, 2013; Taormina and Chau, 2015). In these studies, it was found that the ANN model performed better for short-interval prediction and produced similar results to classical time series models for long-interval forecast. GP provides another possible intelligent solution for nonlinear ocean wave prediction. Real-time wave forecasting at two locations in the Gulf of Mexico (Gaur and Deo, 2008) suggested that GP performed better than ANN for higher interval forecasts. Even the intelligent-technique-based nonlinear models may perform well in handling nonlinearity, they may not be capable of modeling non-stationary data without preprocessing (Cannas *et al.*, 2006; Deka and Prahlada, 2012), especially for long-interval forecast. In addition, a substantial sample size is strictly required in training ANN models, which may imply that a high computational cost is incurred. For example, Deo *et al.* (2001), Agrawal and Deo (2002), Mandal and Prabaharan (2010), and Kamranzad *et al*. (2011) used 80% of the data to train ANN models.

Modeling a nonlinear and non-stationary 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 important features. Hybrid models that combine pre-techniques with single models provide more effective modeling. Time series of wave height are frequently decomposed into several simple components. Then, each component is modeled using single prediction models.

Conventionally, Fourier transform and wavelet analysis are the approaches adopted most often. As Fourier transform is a linear and stationary method, it is unsuitable for nonlinear and non-stationary time series. Wavelet-based models, such as wavelet fuzzy logic model (Özger, 2010) and wavelet neural network (WLNN) (Deka and Prahlada, 2012), have been used for wave forecasting in which the wavelet technique is effective in handling non-stationarity. However, wavelet-based hybrid models have deficiencies in nonlinear and non-stationary wave forecasting. Essentially, wavelet transform is a linear and non-stationary technique. It represents a signal by a linear combination of wavelet basis functions. Its decomposition results for nonlinear data can be misleading (Huang and Wu, 2008; Kim *et al.*, 2012). Furthermore, wavelet analysis suffers from its non-adaptive nature as it applies the same type of basis functions to the entire range of data. A set of basis functions that reflects the time-varying property of a signal is required.

A data-driven technique known as empirical mode decomposition (EMD) has been proposed by Huang *et al.* (1998). This approach is powerful and adaptive in analyzing nonlinear and non-stationary data sets. It provides an effective approach for decomposing a signal into a collection of so-called intrinsic mode functions (IMFs), which can be treated as empirical basis functions. EMD technique acts essentially as a dyadic filter (Flandrin *et al.*, 2004) that separates a complex signal with wide frequency band into relatively simple components with various time scales.

The EMD technique has been widely applied to improve the performance of prediction models (Duan *et al*., 2015; Wang *et al.*, 2015). In this study, a hybrid EMD-AR model was developed for the nonlinear and non-stationary wave forecasting. Compared with the intelligent-technique-based nonlinear models, the AR model is promising in practical engineering applications as it is convenient for real-time model identification, is highly adaptive,

and requires a low computational cost (Zhang and Chu, 2005). However, it suffers from the limitations of nonlinearity and non-stationarity. This situation is overcome by hybridizing the EMD technique with the AR model. Implementation of the EMD-AR model for wave forecast consists of three steps. In the first step, the measured wave time series is decomposed into several stationary components called IMFs. In the second step, AR is used to forecast each component. In the final step, the prediction results of all components are aggregated to obtain the expected wave forecasts.

Derived from the EMD technique and AR model, the EMD-AR model is data-driven, highly adaptive, and suitable for nonlinear and non-stationary time series. An investigation of the EMD-AR model for nonlinear and non-stationary wave forecasting was conducted by using wave data sets from three buoys. The buoys were located at various sites and maintained by the National Data Buoy Center (NDBC), USA. For comparison, the AR model was also studied using the same data sets. The results indicate the superiority of the EMD-AR model and the effectiveness of the EMD technique in extending the scope of the AR model in wave forecasting.

In this paper, the theoretical formulations and numerical schemes of the AR and EMD-AR models are presented first, then brief descriptions of the wave data and accuracy measures are given, finally, numerical results using various significant wave height data sets are presented.

2 Theoretical formulations

2.1 AR prediction model

The AR model considers relations among variables of the time sequence; therefore, the present variable can be represented by using the previous time variable. For a given time series $\{x(t), t=1, 2, \ldots, n\}$, the model is formulated as

$$x(t)=\varphi_1 x(t-1)+\varphi_2 x(t-2)+\ldots+\varphi_p x(t-p)+a(t), \quad t=1, 2, \ldots, n, \tag{1}$$

where p is the model order, $\{\varphi_1, \varphi_2, \ldots, \varphi_p\}$ are parameters of the AR model, which are unknown. The variable $\{a(t), t=1, 2, \ldots, n\}$ is zero-mean white noise. Identification of the AR model shown in Eq. (1) involves the selection of model order p and corresponding parameters $\{\varphi_1, \varphi_2, \ldots, \varphi_p\}$.

A variety of algorithms have been developed for estimating the model parameters, of which, least mean squares (LMS), recursive least squares (RLS), and Levinson–Durbin (L-D) algorithms are mostly used. However, LMS algorithms suffer from low convergence speed and eigenvalue spread problems. The use of the RLS algorithm introduces problem that program code for the sliding-window RLS algorithm is complicated to implement, memory intensive, and potentially numerically unstable (Douglas, 1996). Additionally, the determination of forgetting factor is not always adaptive, leading to non-negligible fluctuations in prediction accuracy. Therefore, the L-D algorithm was adopted to estimate the model parameters in this study.

For a given time sequence $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$, the L-D algorithm for an AR model with a order of p consists of the following steps: (1) compute the autocorrelation matrix $\boldsymbol{R}$ with a size of $(p+1)\times(p+1)$ using Eq. (2); (2) set the initial conditions using Eqs. (3) and (4); (3) compute the coefficients of order k using the coefficients of model order $k-1$ based on Eqs. (5)–(8) until k equals the preset order p.

$$r_k = \frac{1}{n-k}\sum_{i=1}^{n} x_i x_{i+k}, \quad k=0, 1, \cdots, p, \tag{2}$$

where r_k denotes the autocorrelation function of the sequence $\{x_1, x_2, \ldots, x_{n-1}, x_n\}$ for a lag k.

$$\varphi_{1,1} = \rho_1, \tag{3}$$

$$\sigma_1 = r_0(1-\rho_1^2), \tag{4}$$

where $\varphi_{1,1}$ is the first-order model parameter, σ_1 is the variance, and ρ_1 is the reflection coefficient as shown in the following equation:

$$\rho_1 = r_1 / r_0, \tag{5}$$

$$\varphi_{i,k} = \begin{cases} \rho_k, & i=k, \\ \varphi_{i,k-1}-\rho_k\varphi_{k-i,k-1}, & i=1, 2, \cdots, k-1, \end{cases} \tag{6}$$

$$\sigma_k^2 = \sigma_{k-1}^2(1-\rho_k^2), \tag{7}$$

where $\varphi_{i,k}$ is the kth order model parameter and σ_k is the corresponding variance. The kth order reflection coefficient ρ_k is formulated as

$$\rho_k = \frac{r_k - \sum_{i=1}^{k} \varphi_{i,k-1} r_{k-i}}{\sigma_{k-1}^2}. \tag{8}$$

Another problem in AR modeling is the selection of an optimal order. In recent decades, numerous criteria have been proposed to determine the AR order of specified time series. Although it has been a long time since they were first proposed, the Akaike information criterion (Akaike, 1974) and Bayesian information criterion (BIC) (Akaike, 1979) are still the most popular approaches. These criteria have been widely used in various principles of engineering, especially in economic studies. Assume that the residual variance representing the measure of fitness of AR(p) to the data is defined as $\hat{\sigma}_{\mathrm{a}}^2(p)$. It can be formulated as

$$\hat{\sigma}_{\mathrm{a}}^2(p) = \frac{1}{N-p} \sum_{t=p+1}^{N} \left(x_t - \sum_{i=1}^{p} \varphi_i x_{t-i} \right)^2. \tag{9}$$

With the definition of the residual variance, order selection criteria of BIC are briefly described as

$$\mathrm{BIC}(p) = \frac{\lg \hat{\sigma}_{\mathrm{a}}^2(p) + (p+1)\lg N}{N}. \tag{10}$$

In this study, the BIC principle is applied in order selection. The model order p_0 leading to the minimum BIC value is chosen as the optimal order. Once the prediction model as presented in Eq. (1) is determined, a k-step-ahead adaptive predictor can be presented as

$$\hat{x}(t+k) = \begin{cases} \sum_{i=1}^{p} \varphi_i x(t+k-i), & k = 1, \\ \sum_{i=1}^{k-1} \varphi_i \hat{x}(t+i)_{N+i} + \sum_{i=k}^{p} \varphi_i x(t+k-i), & k = 2, 3, \cdots, p, \\ \sum_{i=1}^{p} \varphi_i \hat{x}(t+k-i), & k > p, \end{cases} \tag{11}$$

where $\hat{x}(t+k)$ is the prediction of advancing k steps.

2.2 Hybridization process of the EMD-AR model

Decomposition is a critical part of signal processing. Complex signals are frequently decomposed into several simple components and then the information in each component is analyzed to reduce the complexity and enhance interpretability. EMD was proposed by Huang *et al.* (1998), and it is powerful and adaptive in analyzing the nonlinear and non-stationary data sets. It provides an effective approach to decompose a signal into a collection of so-called IMFs, which can be treated as empirical basis functions driven by data. An IMF result from the EMD procedure should satisfy two conditions: (I) the number of extrema and the number of zero-crossings should differ or be equal to 1 and (II) the local average should be zero, i.e., the mean of the upper envelope defined by the local maxima and the lower envelope defined by the local minima should be zero. The first condition is similar to the traditional narrow band requirements for a stationary Gaussian process (Huang *et al.*, 1998). Therefore, the IMF produced through the EMD procedure is stationary.

For a given sequence $x(t)$, implementation schemes of EMD comprise the following steps: (1) identify the local extrema; (2) generate the upper envelope $u(t)$ and the lower envelope $l(t)$ via spline interpolation among all the local maxima and the local minima, respectively, and then obtain the mean envelope: $m(t)=[l(t)+u(t)]/2$; (3) subtract $m(t)$ from the signal $x(t)$ to obtain the IMF candidate, that is $h(t)=x(t)-m(t)$; (4) verify whether $h(t)$ satisfies the conditions for IMFs and do steps (1)–(4) until $h(t)$ is an IMF; (5) get the nth IMF component $\mathrm{imf}_n(t)=h(t)$ (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.

By implementing the presented algorithm, the signal can be decomposed according to the following Eq. (12). As an example, Fig. 1 displays decomposition results of the significant wave height data shown in Fig. 4a, where it can be clearly seen that the complex wave height time series can be represented by several simple components.

$$x(t) = \sum_{i=1}^{n} \mathrm{imf}_i(t) + r(t). \tag{12}$$

When implementing EMD technique in time series prediction problem, the boundary effects should be taken into account. Researchers have proposed certain techniques for processing boundary effects, such as the characteristic wave extending method (Huang *et al.*, 1998), the ratio extension method (Wu and Riemenschneider, 2010), and the mirror image extending method (Zhao and Huang, 2001). Among the various approaches, the symmetric extending method is the most popular. However, extended results from the symmetric extension method are far from satisfactory. Distinct differences always exist between the extended extrema and the real ones. The influence of end effects on the performance of EMD-based models has been examined by Xiong *et al.* (2014) and Huang *et al.* (2015). They found that prediction models for end effect processing lead to more reasonable extended results. In this study, the AR prediction model presented in Section 2.1 was used in the processing of boundary effects.

Time series of ocean waves are a kind of complicated nonlinear and non-stationary signal that consists of different oscillation scales. The multiple oscillation scales cause difficulties for AR models when conducting wave forecasts. The combination of an EMD model with an AR model provides an effective way to improve wave prediction. The procedure of carrying out wave forecast using the hybrid EMD-AR models comprises three steps (Fig. 2). In the first step, the wave height time series is decomposed into a couple of simple and meaningful IMFs and a residual by EMD. In the second step, prediction of decomposed components is performed individually using the AR model. In the final step, the predictions are aggregated to attain the final predictions.

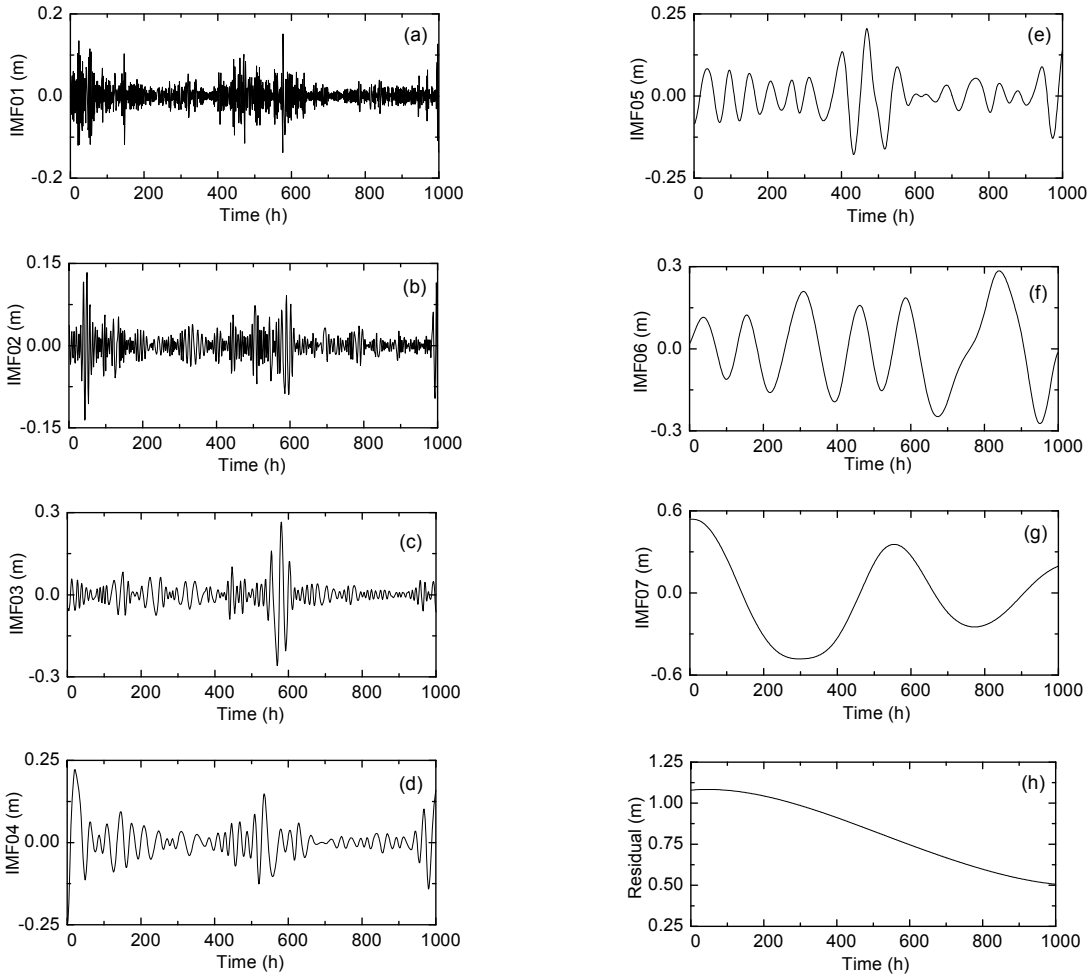

Fig. 1 Decomposition results of significant wave height time series data set I using the EMD technique
Figs. 1a–1h display the simple components with different amplitude and frequency modulations. The data were measured by buoy 42085, which was maintained by the NDBC. Details about the data are provided in Table 1

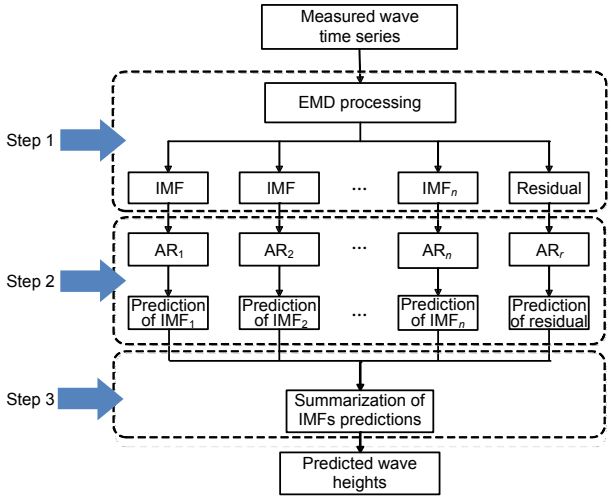

Fig. 2 Implementation of significant wave height forecasting using the EMD-AR model

3 Brief descriptions of the wave data

Ocean wave data from three buoys maintained by the NDBC were used in the forecasting study. The geographical locations where the significant wave height time series are measured and brief non-stationarity analysis of wave data are described in Sections 3.1 and 3.2. Statistical error measures for evaluating prediction performance are presented in Section 3.3.

3.1 Locations and data

To study the performance of the models in forecasting ocean waves with sufficiently different statistical characteristics, significant wave height data measured by buoys on the coast of Ponce (No. 42085), San Juan (No. 41053), and the South Virgin Islands (No. 41052) were chosen. Location information and data availability of these buoys are depicted in Table 1. Some of the hourly time series records (source files from http://www.caricoos.org/drupal/data_download) of the significant wave heights are presented in Fig. 3. The variation in the range of significant wave heights among the three buoys can be seen in Table 1. In view of these differences among the sites, it is reasonable to describe the data from these three buoys as representing a range of geographical and statistical properties (Londhe and Panchang, 2006).

3.2 Non-stationarity analysis

According to traditional definition, a time series, $\{x(t)\}$, is stationary in general, if, for all t,

$$\begin{aligned} &E[x(t)] = \text{constant} < \infty, \\ &E[x^2(t)] < \infty, \\ &E[x(t_1)x(t_2)] = R(t_2 - t_1), \end{aligned} \tag{13}$$

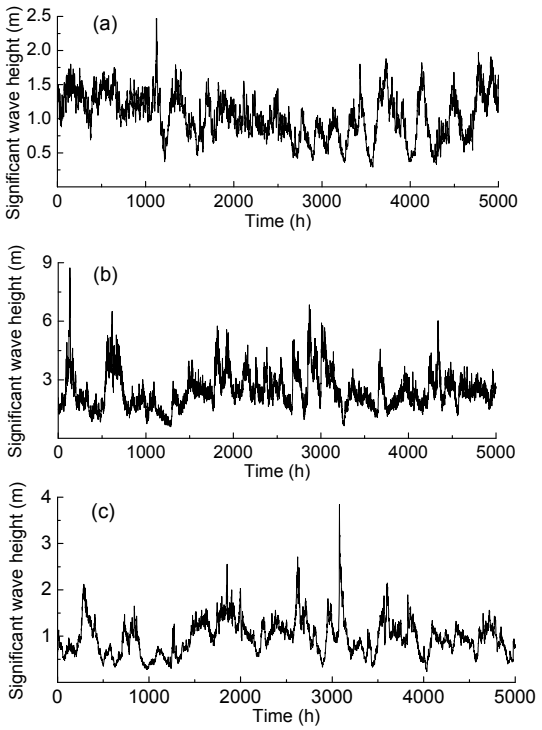

Fig. 3 Significant wave height time series from the wave measurements by buoys: (a) No. 42085, (b) No. 41053, and (c) No. 41052

Table 1 Buoy locations and data availability

Data	Station ID and location	Available data	Range of the significant wave heights (m)
Data set I	No. 42085-Southeast Ponce, Puerto Rico (17°51'36" N, 66°31'25" W)	24/05/2013–26/02/2014	0.25–2.5
Data set II	No. 41053-San Juan, Puerto Rico (18°28'27" N, 66°5'57" W)	23/07/2010–03/05/2013	0.75–9.0
Data set III	No. 41052-South Virgin Islands (18°14'55" N, 64°45'45" W)	15/04/2011–11/05/2013	0.30–4.0

where $E[\cdot]$ is the expected value defined as the ensemble average of the quantity, and R is the covariance function.

Based on the definition of a stationary process, quantitative methods of consecutive statistics are used to analyze the stationarity of significant wave height time series. For stationary time series, their expected value and covariance functions are required to be constants. Fig. 4 shows the expected value and covariance functions of a stationary time series. For specification, the time delay τ in the covariance function $R(\tau)$ is assumed to be 10. It is clearly verified that the expected and covariance functions of the stationary time series are nearly constants. According to the definition of stationary process formulated in Eq. (13), it is demonstrated that the IMF produced by the EMD in Fig. 4a is stationary. Fig. 5 presents the statistical functions of significant wave height data. It shows that expected value functions and covariance functions $R(10)$ are notably time varying, demonstrating the presence of non-stationarity in the significant wave height data.

3.3 Evaluation of forecasting performance

Error measures that are used for the evaluation of forecasting performance usually include the root mean square error (RMSE), the correlation coefficient (r), the scatter index (SI), and the mean absolute error (MAE). Each one of these error criteria has usefulness and limitations (Kalra *et al.*, 2005). For example, the correlation coefficient r is a widely accepted measure of the degree of linear association between the target and the realized outcome, but it is highly sensitive to the extreme values. Hence, they should be viewed together while drawing any inference based on their magnitude.

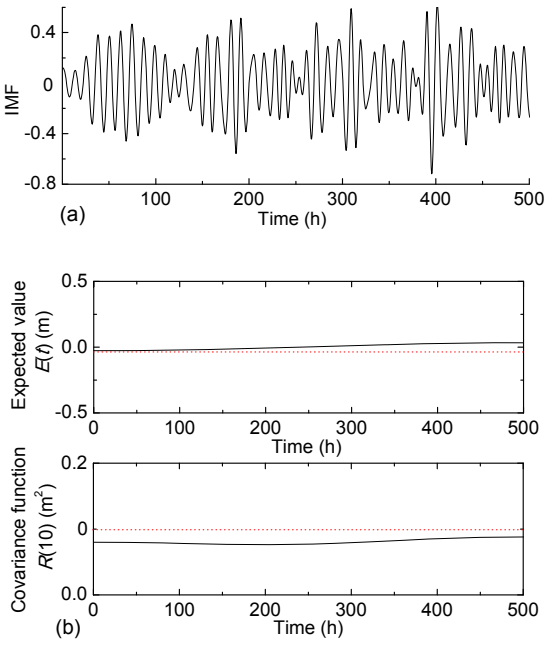

Fig. 4 Expected value and covariance functions of a stationary time sequence
(a) Example of IMF produced by implementing EMD technique; (b) Expected value and covariance functions of IMF

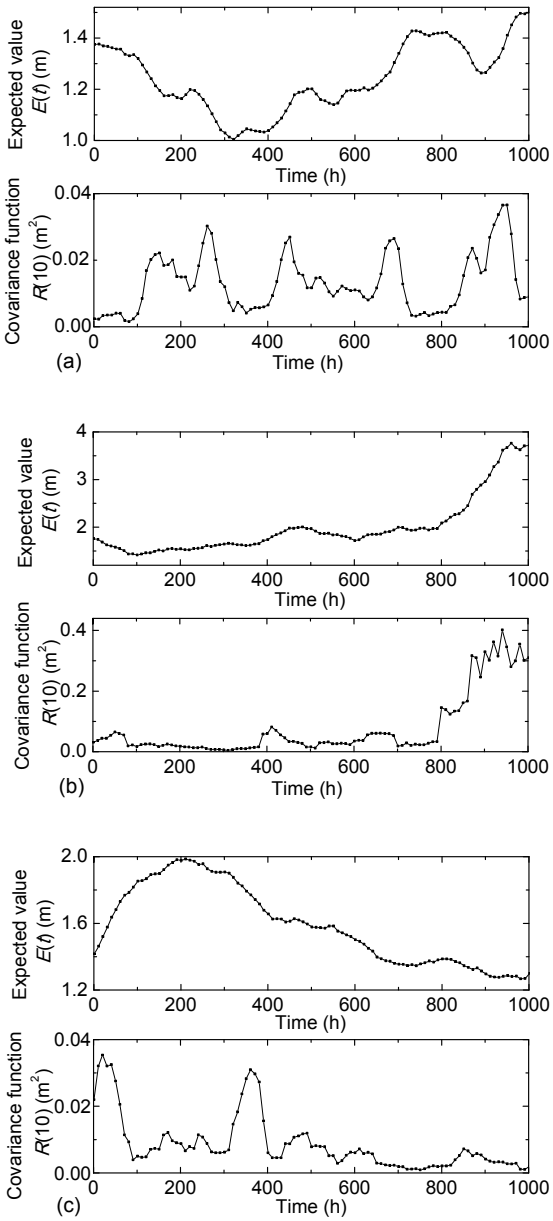

Fig. 5 Statistics of significant wave height data: expected values and covariance functions of data set I (a), data set II (b), and data set III (c)

In this study, prediction results were studied by (I) comparing time histories of the above models' forecasts with measured wave heights, (II) computing the RMSE, the correlation coefficient (r), and the SI as shown in Eqs. (14)–(16), and (III) drawing scatter diagrams and computing the corresponding best-fit line slope. The RMSE is a measure representing the ensemble error of the prediction results. It is proportional to the observed mean. The SI forms a good non-dimensional error measure.

$$r=\frac{\sum_{t=1}^{n}(\hat{x}_t-\hat{x}_m)(x_t-x_m)}{\sqrt{\sum_{t=1}^{n}(\hat{x}_t-\hat{x}_m)^2\sum_{t=1}^{n}(x_t-x_m)^2}}, \tag{14}$$

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_t-x_t)^2}, \tag{15}$$

$$\mathrm{SI}=\frac{\mathrm{RMSE}}{x_m}, \tag{16}$$

where $\hat{x}_t$ is the forecast results with the mean value of $\hat{x}_m$, x_t the measured wave height motions, x_m the mean value of x_t, and n the testing times.

4 Results and discussion

The AR and EMD-AR models were tested using significant wave heights measured by buoys (No. 42085, No. 41053, and No. 41052). A fixed sliding window with a sample size of 500-h wave height records was designed to construct prediction models, while the subsequent 500-h data were used for validation purposes.

4.1 Results

4.1.1 Prediction results using data set I

1-h, 3-h, and 6-h historical predictions of significant wave heights on the coast of Ponce are shown in Figs. 6–8. Scatter diagrams of the forecasts are presented in Figs. 9–11. The values of the error measures, including r, RMSE, and SI, under various lead times are summarized in Table 2 (p.124). Additionally, the error measures of RMSE and r are plotted in Fig. 12 to show the relations between their magnitudes and the prediction lead times.

4.1.2 Prediction results using data set II

Further comparisons of the prediction models were carried out using the significant wave height

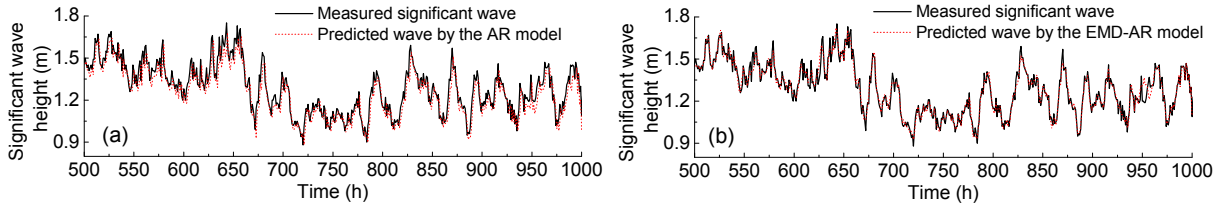

Fig. 6 1-h forecast of significant wave height on the coast of Ponce by AR model (a) and EMD-AR model (b)

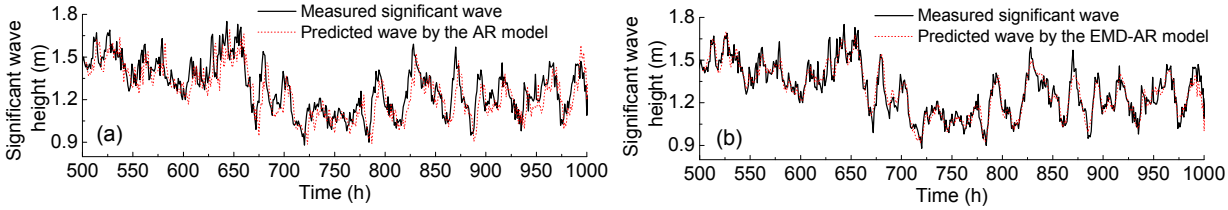

Fig. 7 3-h forecast of significant wave height on the coast of Ponce by AR model (a) and EMD-AR model (b)

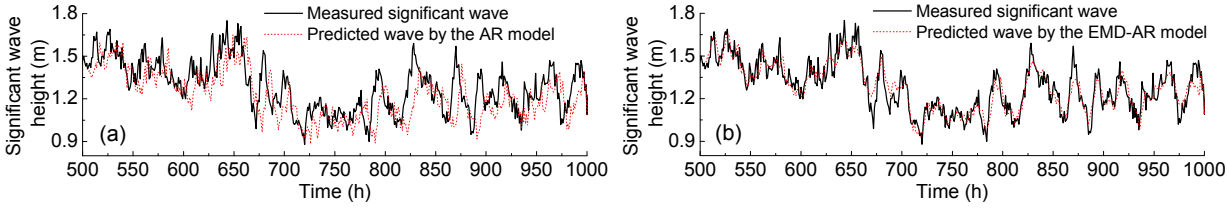

Fig. 8 6-h forecast of significant wave height on the coast of Ponce by AR model (a) and EMD-AR model (b)

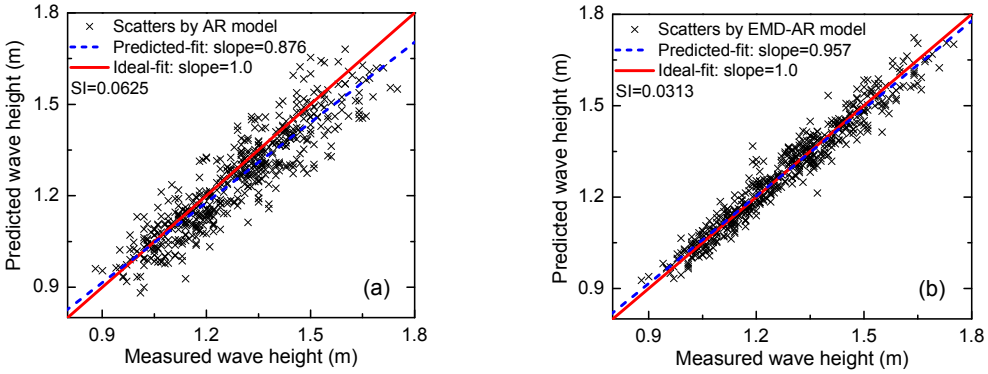

Fig. 9 Scatter diagram of observations and 1-h predictions by AR (a) and EMD-AR (b) models using data set I

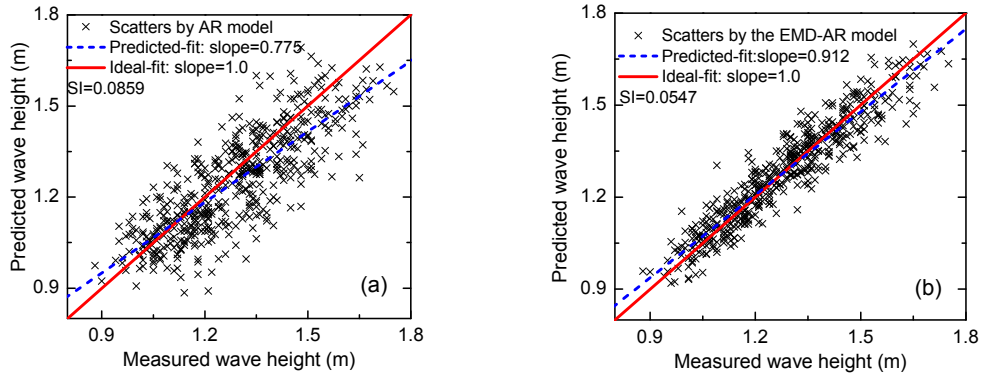

Fig. 10 Scatter diagram of observations and 3-h predictions by AR (a) and EMD-AR (b) models using data set I

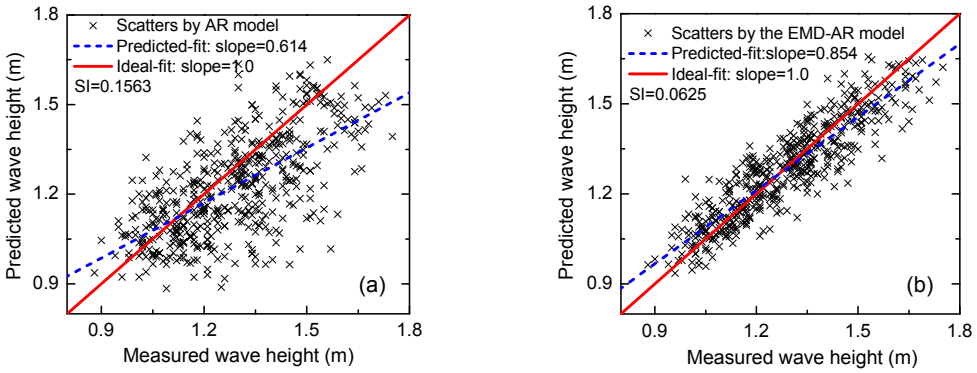

Fig. 11 Scatter diagram of observations and 6-h predictions by AR (a) and EMD-AR (b) models using data set I

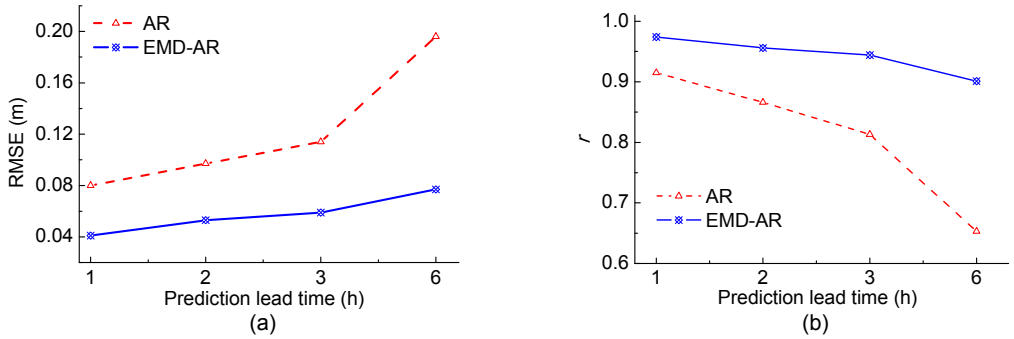

Fig. 12 RMSE (a) and correlation coefficient (b) of prediction models with various lead times using data set I

records measured by buoy 41053 on the coast of San Juan. Figs. 13 and 14 show the 1-h and 6-h predicted time histories, respectively, while Figs. 15 and 16 exhibit the corresponding scatter diagrams. Ensemble error measures are summarized in Table 3, and the RMSE and *r* are plotted in Fig. 17.

4.1.3 Prediction results using data set III

Explorations of the prediction models were consolidated by forecasting simulations using significant wave heights measured by buoy 41052 arranged on the coast of the South Virgin Islands. Similarly, results are represented in the form of historical

Table 2 Error measures of AR and EMD-AR models in predicting significant wave heights in Ponce

Prediction lead time (h)	AR			EMD-AR		
	RMSE	SI	*r*	RMSE	SI	*r*
1	0.08	0.0625	0.92	0.04	0.0313	0.97
2	0.10	0.0781	0.87	0.05	0.0391	0.96
3	0.11	0.0859	0.81	0.07	0.0547	0.94
6	0.20	0.1563	0.65	0.08	0.0625	0.90

Table 3 Error measures of the AR and EMD-AR models in predicting significant wave heights in San Juan

Prediction lead time (h)	AR			EMD-AR		
	RMSE	SI	*r*	RMSE	SI	*r*
1	0.47	0.1794	0.93	0.24	0.0916	0.98
2	0.51	0.1966	0.91	0.30	0.1145	0.96
3	0.52	0.1985	0.90	0.33	0.1256	0.96
6	0.63	0.2405	0.86	0.38	0.1450	0.94

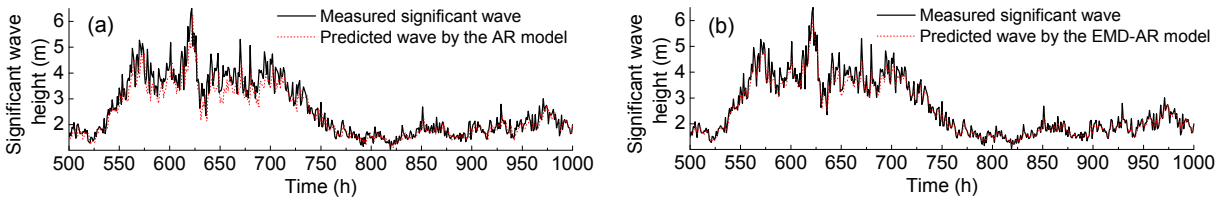

Fig. 13 1-h forecast of significant wave on the coast of San Juan by AR (a) and EMD-AR (b) models

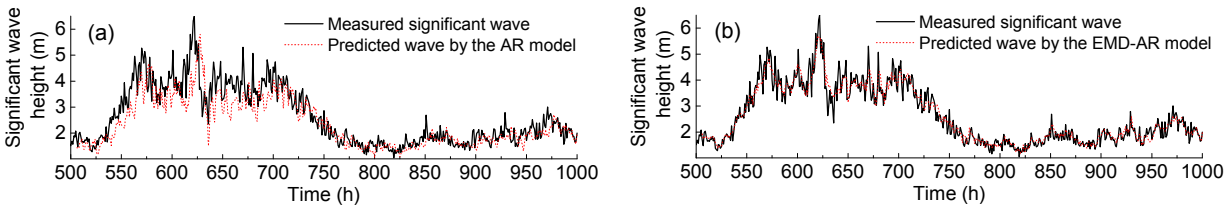

Fig. 14 6-h forecast of significant wave on the coast of San Juan by AR (a) and EMD-AR (b) models

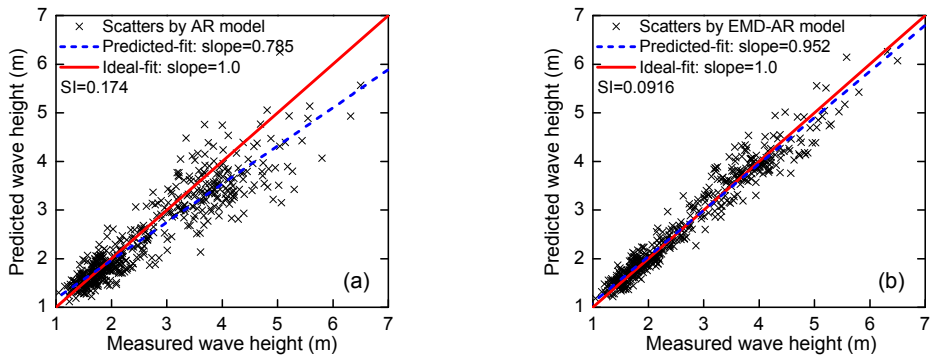

Fig. 15 Scatter diagram of observations and 1-h predictions by AR (a) and EMD-AR (b) models using data set II

time series, scatter diagrams, and error measures. For brevity, only 6-h forecasting time historical results (Fig. 18) and the corresponding scatter diagrams (Fig. 19) are presented. Summaries of error measures used in various prediction lead times are shown in Table 4 and Fig. 20.

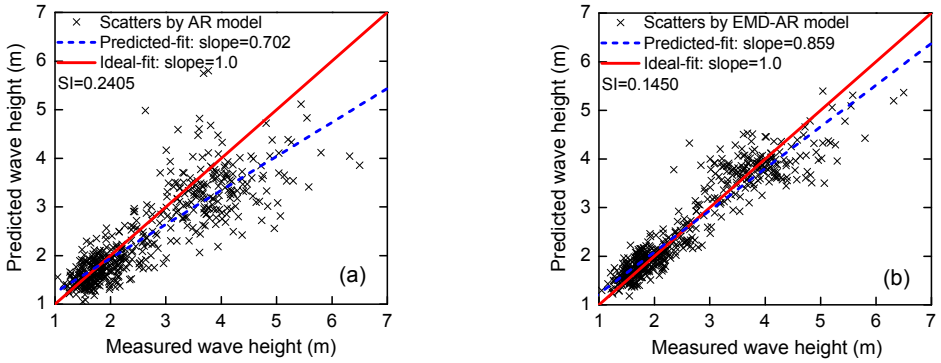

Fig. 16 Scatter diagram of observations and 6-h predictions by AR (a) and EMD-AR (b) models using data set II

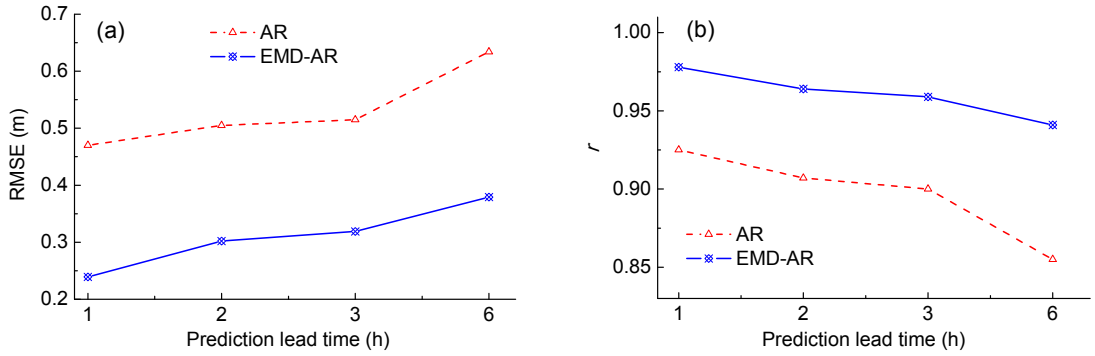

Fig. 17 RMSE (a) and correlation coefficient (b) of prediction models with various lead times using data set II

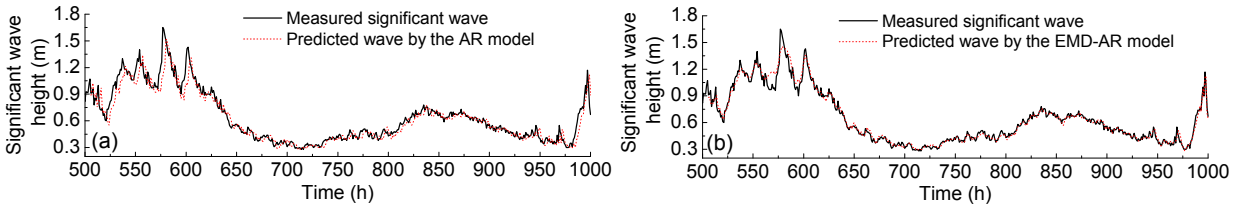

Fig. 18 6-h forecast of significant wave on the coast of South Virgin Islands by AR (a) and EMD-AR (b) models

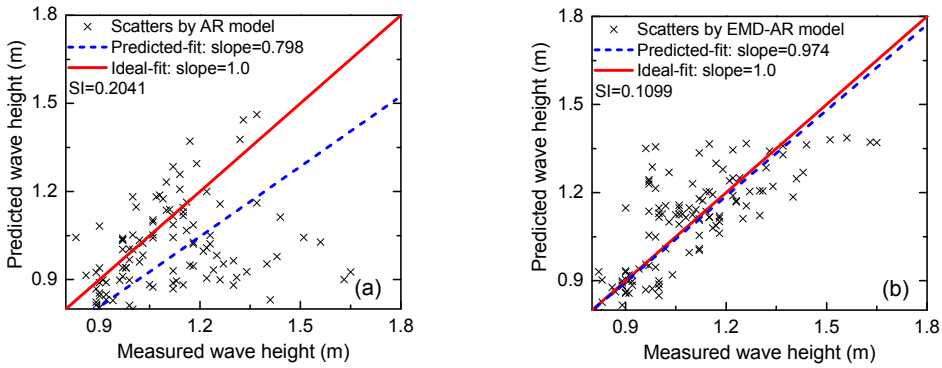

Fig. 19 Scatter diagram of observations and 6-h predictions by AR (a) and EMD-AR (b) models using data set III

Table 4 Error measures of the AR and EMD-AR models in predicting significant wave heights in South Virgin Islands

Prediction lead time (h)	AR			EMD-AR		
	RMSE	SI	*r*	RMSE	SI	*r*
1	0.06	0.0942	0.97	0.03	0.0471	1.00
2	0.07	0.1099	0.98	0.04	0.0628	0.99
3	0.09	0.1413	0.96	0.05	0.0785	0.98
6	0.13	0.2041	0.91	0.07	0.1099	0.97

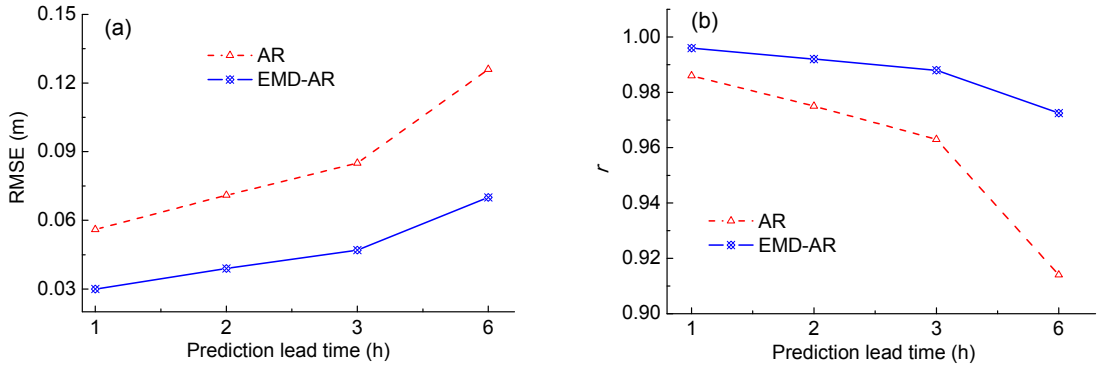

Fig. 20 RMSE (a) and correlation coefficient (b) of prediction models with various lead times using data set III

4.2 Discussion

It is clear from the results that 1-h wave forecasts at various locations using the AR model agree with the measurements to a reasonable degree. As Figs. 6 and 13 suggest, the general patterns of the recorded significant wave height variation in different locations were well captured by the AR model. Tables 2–4 present the values of the forecasting measure errors, where the correlation coefficients for the wave forecasts in Ponce, San Juan, and the South Virgin Islands were 0.92, 0.93, and 0.97, respectively, indicating a relatively high degree of linear association between the predicted and recorded wave heights.

However, prediction errors remain noticeable in the predicted time series as shown in Figs. 6 and 13. Spatial offsets appear as large parts of the troughs and peaks are underestimated. Figs. 9 and 15 show that the best-fit line slopes for the scatters with respect to wave forecasts in Ponce and San Juan are only 0.857 and 0.785, respectively. In addition to the spatial offsets, Figs. 6 and 13 imply that even if the peaks and troughs were well predicted by the AR model, a shift between the recorded and predicted wave time series can still be noted. The shift is a kind of prediction error that can also be found in other research studies of wave forecasting using the AR model (Deo and Sridhar, 1998) and ANN (Londhe and Panchang, 2006). The shift results mainly from the non-stationarity hidden in the measured wave time series. Even if the nonlinear ANN is used to forecast the nonlinear and non-stationary wave height, the shift remains.

The shift is proportional to the lead time. In Figs. 6–8, it is easy to see that the shift between the AR-based predicted and recorded wave time series increases as the lead time grows. Predictions by the AR model in San Juan and the South Virgin Islands support these observations. As presented in Tables 2–4 and in Figs. 12, 17, and 20, the RMSE and SI increase, while the correlation coefficient decreases with the increase of the lead time.

Owing to the linear and stationary limitations, the AR model fails to predict the nonlinear and non-stationary wave heights accurately when the lead time reaches 6 h. The best-fit line slopes of the scatters with respect to wave forecasts in Ponce, San Juan, and the South Virgin Islands are only 0.614, 0.702, and 0.798, respectively, indicating a relatively low level of forecasting accuracy. The nonlinear and

non-stationary wave forecasts are considerably improved by using the proposed EMD-AR model. The predictions of the hybrid EMD-AR show better agreement with the targets. When the lead time is short, not only are the peaks and troughs of the targets precisely captured for the most part but also the short-term fluctuations in the sequence are reproduced remarkably well (Figs. 6 and 13). For instance, the spatial offsets resulting from the observations and the predictions by the AR model are quite noticeable in the range of 650–700 h, especially for forecasts with a large lead time (Fig. 14). This situation is noticeably improved by introducing the EMD technique (Figs. 13 and 14). Additionally, the best-fit line slopes in the scatter of 1-h wave forecasts using the AR and EMD-AR models in Ponce (Fig. 9) are 0.876 and 0.957, respectively. Despite the small spatial offsets relative to the target when the lead time grows, forecasts of the EMD-AR model display a level of fidelity in the measured significant wave heights which is certainly acceptable for most practical applications.

As shown in Figs. 6–8, 13, 14, and 18, the shifts between the predicted and recorded wave time series in various locations were eliminated by using the EMD-AR model instead of the AR model. This improvement is confirmed by comparing the error measures of the EMD-AR and AR models. For example, the correlation coefficients of the 6-h predictions in Ponce, San Juan, and the South Virgin Islands by the AR model were 0.65, 0.86, and 0.91, respectively, while those by the EMD-AR model are 0.90, 0.94, and 0.97, respectively. Meanwhile, the SIs of the 6-h predictions in these locations by the AR model were 0.1563, 0.2405, and 0.2041, while those of the EMD-AR model were 0.0625, 0.1450, and 0.1099, respectively. In addition, Tables 2–4 and Figs. 12, 17, and 20 summarize error measures with various lead times, providing general evidence for the above claims. The EMD-AR model led to lower ensemble RMSE and higher r. The graphs in Figs. 12, 17, and 20 combined with Tables 2–4 demonstrate large improvements in prediction accuracy by using the EMD technique in the AR model. Considerable reductions in RMSE and increases in correlation coefficient were obtained. Taking 6-h wave forecasts as an example, in Table 2, the reduction in RMSE was about 60%, while the increase in the correlation coefficient was more than 50%.

5 Concluding remarks

This study developed a hybrid EMD-AR model to improve the accuracy of prediction of nonlinear and non-stationary waves. The EMD-AR and AR models were compared using wave data with various geographical and statistical properties measured by NDBC buoys in Ponce, San Juan, and the South Virgin Islands. Consistent results were obtained from the predictions of significant wave heights in different locations. For short-interval predictions, the AR model may produce reasonable results. However, spatial offsets and shifts occur widely in the nonlinear and non-stationary wave forecast. This is because the AR model is suitable only for linear and stationary time series prediction, whereas nonlinearity and non-stationarity are features of all the measurements. These errors increased as the lead time grew. This difficulty was overcome by the proposed hybrid EMD-AR model. Owing to the capability of the EMD technique in processing nonlinearity and non-stationarity, the accuracy of the wave forecast was greatly improved. Not only were the general tendencies satisfactorily reproduced but also most part of the peaks and troughs were correctly captured. Considerable improvements in prediction accuracy were obtained using the hybrid EMD-AR model. Graphs related to predicted time histories (Figs. 6–8, 13, 14, and 18) suggest that the shifts between the predicted and recorded wave time series were eliminated by the EMD technique. The superiority of the EMD-AR model to the AR model was confirmed by the ensemble of the smaller RMSE and SI, and higher r. However, the hybrid EMD-AR model has a limitation: it requires more computational resource than the single AR model.

References

Agrawal, J.D., Deo, M.C., 2002. On-line wave prediction. *Marine Structures*, **15**(1):57-74. http://dx.doi.org/10.1016/S0951-8339(01)00014-4

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6): 716-723.

http://dx.doi.org/10.1109/TAC.1974.1100705

Akaike, H., 1979. A Bayesian extension of the minimum AIC procedure of autoregressive model fitting. *Biometrika*, **66**(2):237-242. http://dx.doi.org/10.1093/biomet/66.2.237

Cannas, B., Fanni, A., See, L., *et al*., 2006. Data preprocessing for river flow forecasting using neural networks: wavelet transforms and data partitioning. *Physics and Chemistry of the Earth, Parts A/B/C*, **31**(18):1164-1171. http://dx.doi.org/10.1016/j.pce.2006.03.020

Chau, K.W., 2007. Application of a PSO-based neural network in analysis of outcomes of construction claims. *Automation in Construction*, **16**(5):642-646. http://dx.doi.org/10.1016/j.autcon.2006.11.008

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

Deo, M.C., Sridhar, N.C., 1998. Real time wave forecasting using neural networks. *Ocean Engineering*, **26**(3):191-203. http://dx.doi.org/10.1016/S0029-8018(97)10025-7

Deo, M.C., Jha, A., Chaphekar, A.S., *et al*., 2001. Neural network for wave forecasting. *Ocean Engineering*, **28**(7):889-898. http://dx.doi.org/10.1016/S0029-8018(00)00027-5

Douglas, S.C., 1996. Efficient approximate implementations of the fast affine projection algorithm using orthogonal transforms. IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, USA, **3**:1656-1659. http://dx.doi.org/10.1109/ICASSP.1996.544123

Duan, W.Y., Huang, L.M., Han, Y., *et al.*, 2015. A hybrid AR-EMD-SVR model for the short-term forecast of nonlinear and non-stationary ship motion. *Journal of Zhejiang University-SCIENCE A* (*Applied Physics & Engineering*), **16**(7):562-576. http://dx.doi.org/10.1631/jzus.A1500040

Engle, R.F., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica*, **50**(4):987-1008. http://dx.doi.org/10.2307/1912773

Flandrin, P., Rilling, G., Gonçalvés, P., 2004. Empirical mode decomposition as a filter bank. *IEEE Signal Processing Letters*, **11**(2):112-114. http://dx.doi.org/10.1109/LSP.2003.821662

Gaur, S., Deo, M.C., 2008. Real-time wave forecasting using genetic programming. *Ocean Engineering*, **35**(11-12):1166-1172. http://dx.doi.org/10.1016/j.oceaneng.2008.04.007

Hannan, E.J., 1982. A note on bilinear time series models. *Stochastic Processes and Their Applications*, **12**(2):221-224 http://dx.doi.org/10.1016/0304-4149(82)90044-8

Huang, L.M., Duan, W.Y., Han, Y., *et al*., 2015. Extending the scope of AR model in forecasting non-stationary ship motion by using AR-EMD technique. *Journal of Ship Mechanics*, **19**(9):1033-1049 (in Chinese). http://dx.doi.org/10.3969/j.issn.1007-7294.2015.09.002

Huang, N.E., Wu, Z.H., 2008. A review on Hilbert-Huang transform: method and its applications to geophysical studies. *Reviews of Geophysics*, **46**(2):2007RG000228. http://dx.doi.org/10.1029/2007RG000228

Huang, N.E., Shen, Z., Long, S.R., *et al*., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **454**(1971):903-995. http://dx.doi.org/10.1098/rspa.1998.0193

Jain, P., Deo, M.C., 2007. Real-time wave forecasts off the western Indian coast. *Applied Ocean Research*, **29**(1-2):72-79. http://dx.doi.org/10.1016/j.apor.2007.05.003

Janssen, P.A.E.M., 2008. Progress in ocean wave forecasting. *Journal of Computational Physics*, **227**(7):3572-3594. http://dx.doi.org/10.1016/j.jcp.2007.04.029

Kalra, R., Deo, M.C., Kumar, R., *et al*., 2005. RBF network for spatial mapping of wave heights. *Marine Structures*, **18**(3):289-300. http://dx.doi.org/10.1016/j.marstruc.2005.09.003

Kamranzad, B., Etemad-Shahidi, A., Kazeminezhad, M.H., 2011. Wave height forecasting in Dayyer, the Persian Gulf. *Ocean Engineering*, **38**(1):248-255. http://dx.doi.org/10.1016/j.oceaneng.2010.10.004

Kim, D., Kim, K.O., Oh, H.S., 2012. Extending the scope of empirical mode decomposition by smoothing. *EURASIP Journal on Advances in Signal Processing*, **2012**(1):168. http://dx.doi.org/10.1186/1687-6180-2012-168

Komen, G.J., Cavaleri, L., Donelan, M., *et al*., 1994. Dynamics and Modelling of Ocean Waves. Cambridge University Press, Cambridge. http://dx.doi.org/10.1017/CBO9780511628955

Li, G., Weiss, G., Mueller, M., *et al*., 2012. Wave energy converter control by wave prediction and dynamic programming. *Renewable Energy*, **48**:392-403. http://dx.doi.org/10.1016/j.renene.2012.05.003

Londhe, S.N., Panchang, V., 2006. One-day wave forecasts based on artificial neural networks. *Journal of Atmospheric and Oceanic Technology*, **23**(11):1593-1603. http://dx.doi.org/10.1175/JTECH1932.1

Mandal, S., Prabaharan, N., 2010. Ocean wave prediction using numerical and neural network models. *The Open Ocean Engineering Journal*, **3**(1):12-17. http://dx.doi.org/10.2174/1874835X01003010012

Özger, M., 2010. Significant wave height forecasting using wavelet fuzzy logic approach. *Ocean Engineering*, **37**(16):1443-1451. http://dx.doi.org/10.1016/j.oceaneng.2010.07.009

Sandhya, K.G., Balakrishnan Nair, T.M., Bhaskaran, P.K., *et al*., 2014. Wave forecasting system for operational use and its validation at coastal Puducherry, east coast of India. *Ocean Engineering*, **80**:64-72.

http://dx.doi.org/10.1016/j.oceaneng.2014.01.009

Taormina, R., Chau, K.W., 2015. Neural network river forecasting with multi-objective fully informed particle swarm optimization. *Journal of Hydroinformatics*, **17**(1):99-113. http://dx.doi.org/10.2166/hydro.2014.116

The Wamdi Group, 1988. The WAM model—a third generation ocean wave prediction model. *Journal of Physical Oceanography*, **18**(12):1775-1810. http://dx.doi.org/10.1175/1520-0485(1988)018<1775:TWMTGO>2.0.CO;2

Tolman, H.L., 2014. User Manual and System Documentation of WAVEWATCH III® Version 4.18. Tech. Note 316, NOAA/NWS/NCEP/MMAB, College Park, MD, USA, p.282.

Tong, H., Lim, K.S., 1980. Threshold autoregressive, limit cycles and cyclical data. *Journal of the Royal Statistical Society Series B*, **42**(3):245-292.

Wang, W.C., Chau, K.W., Xu, D.M., *et al*., 2015. Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. *Water Resources Management*, **29**(8):2655-2675. http://dx.doi.org/10.1007/s11269-015-0962-6

Wu, C.L., Chau, K.W., 2013. Prediction of rainfall time series using modular soft computing methods. *Engineering Applications of Artificial Intelligence*, **26**(3):997-1007. http://dx.doi.org/10.1016/j.engappai.2012.05.023

Wu, Q., Riemenschneider, S.D., 2010. Boundary extension and stop criteria for empirical mode decomposition. *Advances in Adaptive Data Analysis*, **02**(02):157-169. http://dx.doi.org/10.1142/S1793536910000434

Xiong, T., Bao, Y.K., Hu, Z.Y., 2014. Does restraining end effect matter in EMD-based modeling framework for time series prediction? Some experimental evidences. *Neurocomputing*, **123**:174-184. http://dx.doi.org/10.1016/j.neucom.2013.07.004

Zhang, G.Q., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with artificial neural networks: the state of art. *International Journal of Forecasting*, **14**(1):35-62. http://dx.doi.org/10.1016/S0169-2070(97)00044-7

Zhang, J., Chu, F., 2005. Real-time modeling and prediction of physiological hand tremor. IEEE International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, USA, **5**:v/645-v/648. http://dx.doi.org/10.1109/ICASSP.2005.1416386

Zhao, J.P., Huang, D.J., 2001. Mirror extending and circular spline function for empirical mode decomposition method. *Journal of Zhejiang University-SCIENCE*, **2**(3):247-252. http://dx.doi.org/10.1631/jzus.2001.0247

中文概要

题　目：一种用于非线性非平稳波浪极短期预报的复合经验模态分解自回归模型

目　的：相对于由能量平衡方程得到的数值预报模型和以神经网络为代表的非线性模型而言，自回归（AR）模型在波浪预报中具有计算效率高、自适应性强和建模所需的样本小等优点，但同时存在局限于平稳线性假设的缺陷。针对非线性非平稳波浪的极短期预报问题，提出一种复合的经验模态分解自回归预报模型，提高波浪预报精度。

创新点：1. 研究非线性非平稳波浪极短期预报问题，提出一种复合的预报方法；2. 基于三个不同地理位置的海洋波浪实测数据对预测模型进行验证，并分析非线性非平稳性对波浪预报结果的影响。

方　法：1. 在 AR 模型中引入经验模态分解（EMD）方法，形成复合的 EMD-AR 预报模型；2. 分析实测波浪数据的非线性和非平稳性特点，并基于实测波浪数据获得 AR 模型和 EMD-AR 模型的预报结果；3. 基于多种预报误差度量分析 AR 模型和 EMD-AR 模型的预报性能以及非线性非平稳性对波浪预报结果的影响。

结　论：1. 波浪非线性和非平稳性会导致 AR 预报模型精度降低。预报误差中，幅值上的偏差主要由波浪的非线性引起，而相位上的偏差则是源于波浪的非平稳性；2. EMD 方法能够有效地克服波浪非线性和非平稳性对 AR 模型在精度上所带来的不良影响，在精度上 EMD-AR 模型的预报结果较 AR 模型有较大提高。

关键词：波浪预报；非线性和非平稳性；自回归模型；经验模态分解；经验模态分解自回归模型