Knowledge-Based Systems 125 (2017) 39–52

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Red tide time series forecasting by combining ARIMA and deep belief network

Mengjiao Qin[a], Zhihang Li[b], Zhenhong Du[a,c,*]

[a] School of Earth Sciences, Zhejiang University, Hangzhou 310027, China
[b] Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CAS), Beijing 100190, China
[c] Zhejiang Provincial Key Laboratory of Geographic Information Science, Hangzhou 310028, China

ARTICLE INFO

Article history:
Received 2 August 2016
Revised 29 March 2017
Accepted 30 March 2017
Available online 5 April 2017

Keywords:
Red tide forecasting
ARIMA
DBN
PSO
ARIMA-DBN

ABSTRACT

The red tide occurs frequently in recent years. The process of the growth, reproduction, extinction of the red tide algal has a complex nonlinear relationship with the environmental factors. The environmental factors have characteristics including time continuity and spatial heterogeneity. These characteristics make it arduous to forecast red tide. This paper mainly analyzes the related factors of the red tide disasters. Based on the strong forecasting ability of Autoregressive Integrated Moving Average (ARIMA) model and the powerful expression ability of Deep Belief Network (DBN) on nonlinear relationships, a hybrid model which combines ARIMA and DBN is proposed for red tide forecasting. The corresponding ARIMA model is built for each environmental factor in different coastal areas to describe the temporal correlation and spatial heterogeneity. The DBN serves to capture the complex nonlinear relationship between the environmental factors and the red tide biomass, and then realizes the warning of red tide in advance. Furthermore, Particle swarm optimization (PSO) is introduced to enhance the speed of model training. Finally, ship monitoring data collected in Zhoushan coastal area and Wenzhou coastal area during 2008–2014 is used as the experimental dataset. The proposed ARIMA-DBN model is applied to forecasting red tide. The experimental results demonstrate that the proposed method achieves a good forecast of red tide.

1. Introduction

Red tide is a temporary natural phenomenon involving harmful algal blooms (HABs) in company with a changing sea color from normal to red or reddish brown, which has a bad influence on coast environment and sea ecosystems [1]. The occurrence of red tide is the result of the mixed effects of biological, chemical, hydrological and meteorological factors. These factors lead to the seasonal, spatial heterogeneity and complex nonlinear relationship, bringing great challenge to the prediction of red tide in the field of Marine Science.

Along with the growth of population and the rapid development of economy, the burden of the marine environment is increasing. Red tide has become one of the worst marine disasters effecting on ecological environment, which results from the large amount of untreated waste water directly discharged into the ocean. The frequent occurrence of red tide damages the marine ecological balance, which causes huge economic losses to the marine fishery and the related resources.

In recent years, more attention is attached to the marine ecological environment protection in the coastal area. The monitoring method of marine ecological environment is becoming more and more diversified, such as marine station, survey monitoring ship, satellite remote sensing and buoy. Over the years, the volume of the obtained data for red tide is growing rapidly. Because the monitoring contents contain features of time and space, the data with large scale, high dimension and spatial heterogeneity can be derived. Hence, one burning issue is how to use appropriate methods to analyze red tide data and extract valuable information, so as to reveal the law of occurrence of red tide disaster and forecast the red tide in the future.

Existing red tide prediction methods include classification methods (eg. Decision Tree (DT) [34], fuzzy c-means (FCM) [2]), regression methods (eg. Back Prorogation (BP) neural network [35], Radial Basis Function (RBF) neural network [4]). These two kinds of methods can just predict the red tide in the current time. They are not able to forecast the future red tide before occurrence, which is of great importance in the early-warning of red tide disasters.

* Corresponding author.
E-mail addresses: qinmengjiao@zju.edu.cn, qin_mengjiao@163.com (M. Qin), zhihang.li@cripac.ia.ac.cn (Z. Li), duzhenhong@zju.edu.cn (Z. Du).

http://dx.doi.org/10.1016/j.knosys.2017.03.027
0950-7051/

It is obvious that time series forecasting is indispensable in the urgent issue. However, classic methods of time series forecasting such as ARIMA [5] can only model linear patterns in time series. Though the widely used hybrid model ARIMA-BP [18] has solved the nonlinear patterns, it is extremely unstable as the network layer goes deeper. Moreover, it is easy to fall into local minimum, which has a marked impact on the accuracy of the results.

In this paper, a novel ARIMA-DBN hybrid model is proposed to predict the possibility of occurrence and the future trends of red tide. ARIMA Model describes the temporal correlation and spatial heterogeneity of environmental factors, while DBN serves to capture the complex nonlinear relationship between the environmental factors and the red tide biomass. Combining the advantages of DBN and ARIMA, ARIMA-DBM not only models the time series, but also extracts the rule of different features. In addition, ARIMA-DBM is trained in a greedy manner, i.e. layer-wise pre-training, which permits training deeper network and alleviate trapping into local minimum [21,27]. Therefore, benefiting from deeper layers, ARIMA-DBM has a desirable stability and learning ability.

The remainder of this paper is structured as follows: Section 2 gives an overview of works related to this paper, Section 3 describes the models and algorithms for red tide prediction, Section 4 introduces the evaluation criteria and presents the experimental results, and Section 5 discusses the results and makes the conclusion.

2. Related works

2.1. Red tide forecasting

Red tide forecasting has always been a hot research spot since it is closely related to the ecological environment and marine fishery. Due to the diversity of monitoring techniques, large amount of heterogeneous data are available. With remote sensing imagery, [32] forecasted region-wide risk maps and has successfully predicted the observed algal bloom occurrences in Hong Kong waters over the past 20 years; [33] aimed at inferring the red tide trend by analyzing the historical red tide events, seasons and red tide causative species; [40] studied the anticipated linkages between harmful algal blooms and climate change, and attempted to ascertain the earliest signals of harmful algal blooms.

In terms of drifting buoy data and ship monitoring data, the methods can be roughly categorized into two groups: classification methods and regression methods. In the classification methods, red tide prediction is regarded as a classification problem. They divide outputs into several classes which are manually set before prediction. Unlike the classification methods, regression methods like ARIMA-DBN in paper, treat the red tide prediction as a regression problem. The outputs of these methods are continuous values rather than discrete classes.

Many classification methods have obtained good results. Sun Park proposed a red tide prediction approach by integrating ensemble method and fuzzy reasoning algorithm in 2014 [1]. Xiaomei Hu forecasted the red tide using the algorithm of fuzzy c-means (FCM) [2]. The experimental results indicated that the FCM had a strong ability of de-noise in red tide prediction. [3] took the phytoplankton and the biotoxin data and remote sensing imagery into consideration, and finally gained the early warning of harmful algae blooms. Other machine learning methods like Decision Tree (DT) [34] was also applicable in red tide prediction.

Regression methods need to predict continuous values more precisely. Existing approaches include Back Prorogation (BP) neural network [35], Radial Basis Function (RBF) neural network [4], Generalized Regression Neural Network (GRNN) [4] etc. GRNN is a type of shallow network, so its training procedure is easy. However, it is the structure of shallow network that limits its learning capacity. Therefore, GRNN is fit to model the small dataset. On the contrary, BP neural network and RBF neural network work well with large amount of samples, but BP is very sensitive to the initialization and outlier. Since the center of basis function is randomly selected from the data, the performance of RBF is not so satisfactory, but the convergence rate of RBF is faster than BP.

However, most regression approaches are applied to current prediction, which means they can well predict red tide with the current input, but are not able to forecast the future values through the historical data. Future forecasting is the essence of the red tide early-warning and the time series forecasting approaches are the main tool to deal with this problem.

2.2. Time series forecasting

Time series forecasting is to study from the historical observations of given variable, then model its underlying process to forecast the future values. Because of the passion for future exploration, time series forecasting has been employed to various applications, and has achieved great success. It is clearly that time series forecasting is essential to red tide forecasting. More attention ought to be placed to mining the law of red tide time series.

A traditional method named Autoregressive Integrated Moving Average Model (ARIMA) [5] has been applied to various fields, such as financial [6], business [7], short-term traffic flow [8] and carbon price [9]. ARIMA models assume that the present data is a linear function of past data points and past errors. They also assume that the errors are white in nature, and require that the data is stationary before fitting a linear equation [10].

ARIMA can well model linear patterns in time series; however, it is not applicable in nonlinear patterns. With the development of machine learning, many machine learning methods become very important nonlinear techniques in the time series forecasting field. There are lots of examples. Artificial neural network (ANN) was utilized to forecast the wind speed [11]. Different kinds of neuron models were trained for financial time series prediction [12,36]. Support vector machine (SVM) [13] and Bayesian approach [14] were also improved as nonlinear time series forecasting techniques. All these proposed methods have been proved to work well in nonlinear time series forecasting.

In order to combine the advantages of linear and nonlinear approaches, large amount of hybrid approaches [15–20] have been devoted to time series forecasting. Most of the hybrid models consist of ARIMA models and ANN models. In these hybrid models, the linear part of the historical data are handled by ARIMAs whereas the nonlinear part to be processed by the Artificial Intelligent models. In terms of the strong linear forecasting ability of ARIMA and the nonlinear expression ability of Artificial Intelligence, the case studies have achieved good experimental results.

2.3. Deep learning

The concept of deep learning originated from artificial neural network, and was first proposed by Hinton in 2006 [27]. Deep learning means a multi-layer perceptron structure with multiple hidden layers. Once proposed, deep learning caused tremendous repercussions throughout the world.

Deep learning has attracted much attention for its excellent performance in computer vision [21–23], natural language processing [24–26] and representation learning [27]. Multiple levels of nonlinear operations, such as many hidden layers, are able to extract different features from diverse perspectives, which contributes to its powerful non-linear expressive capacity. [21] proposed a new neural network, called AlexNet, applied it to the image classification successfully, and won the first prize in the famous competition ImageNet LSVRC-2010. [24] utilized the

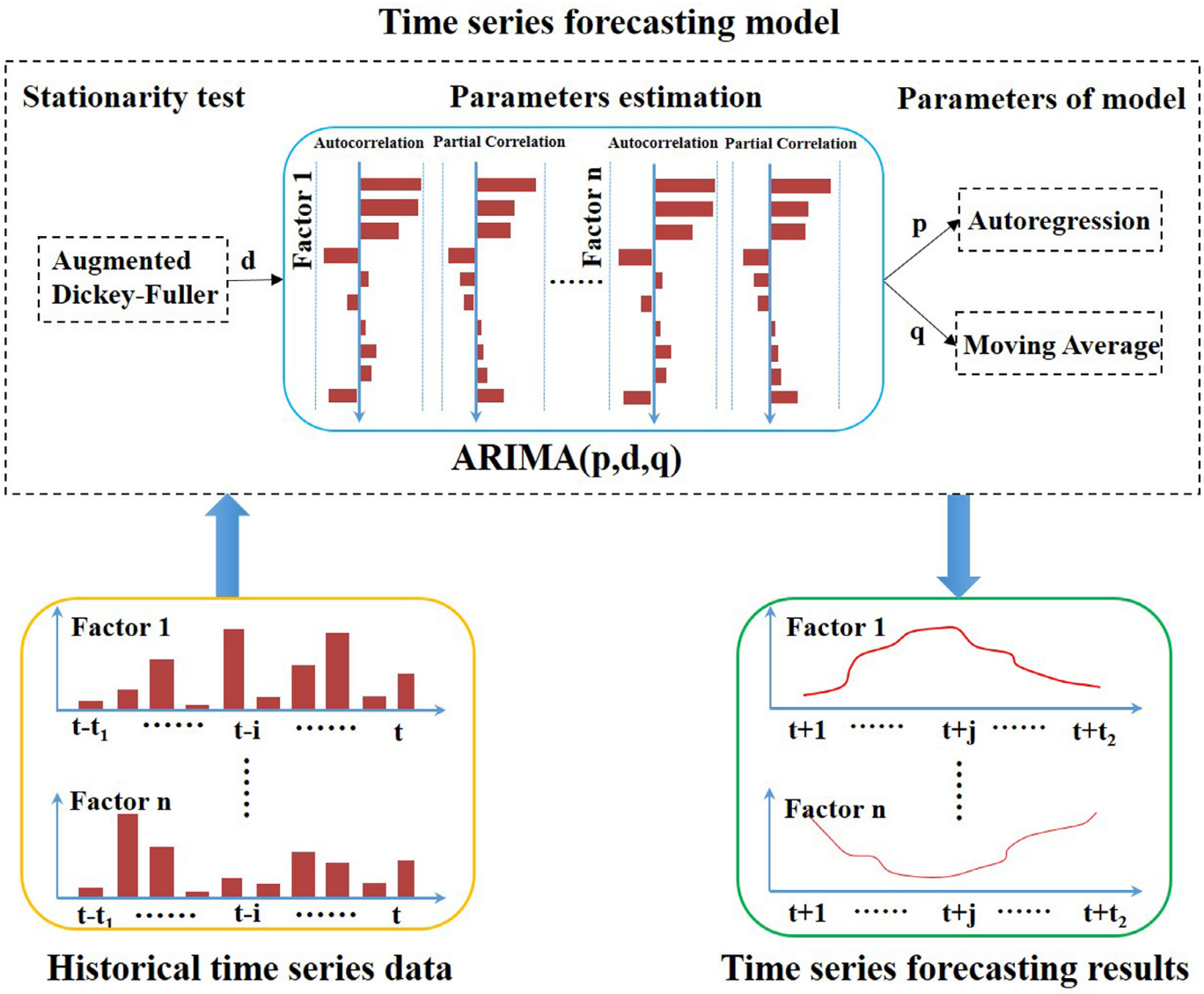

Fig. 1. The flow diagram of ARIMA model: a. Stationarity test (*Step 1*); b. Parameters estimation (*Step 2 & Step 3*); c. Prediction (*Step 4*). ARIMA is applied to forecast the future values of the environmental factors through studying from the historical time series data.

convolutional neural network for language model. Meanwhile, [22] adapted deep learning to obtain a better face feature and achieved a high accuracy. [27] proposed a new fast learning algorithm for deep belief networks, which pre-trained each layer at a time and then fine-tuned the weight on the whole network. Consequently, deep learning has been applied to many fields and achieves an excellent effect in virtue of its generalization and robustness.

It is natural that deep learning has also been involved to forecast time series. [37] improved Deep Belief Network with restricted Boltzmann machines for time series prediction, however, it was not able to forecast the future trends and values of time series. [38] combined unsupervised future learning and deep learning for time-series modeling. This model can work well in stock prediction, speech recognition, motion capture etc. [39] discussed deep learning for both regression problem and time series forecasting problem.

Red tide is an intricate and complex phenomenon, which is caused by a variety of environmental factors. Therefore, we take advantage of the powerful ability of expression of deep learning to discover the complex nonlinear relationship between different factors and compute the red tide biomass.

2.4. Contributions

A novel hybrid model ARIMA-DBN for red tide forecast is proposed in this paper. ARIMA serves to describe the temporal correlation and spatial heterogeneity of each influence factor. DBN contributes to capture the intricate relationship between the factors and the red tide biomass, which reflects the development of red tide.

The innovation points in this paper are:

(1) ARIMA is employed for discovering the law of each environmental factor on the historical data, so as to forecast the future values of the factors. Because the laws of each factor are different in multiple coastal areas, the parameters of the model is not the constant. In this work, ARIMA describes not only the temporal information, but also the spatial heterogeneity.
(2) DBN has a powerful feature expression ability. Since the law of the environmental factors resulting in red tide is seriously com-

plicated, DBN is used to seek for the complex relations between them.

(3) The parameters of ARIMA-DBN are flexible and variable. Therefore, Particle Swarm Optimization (PSO), as an optimization method, is adopted to search the proper solutions and accelerate the training process.

3. Methodology

3.1. ARIMA model

In the ARIMA model, each environmental factor, such as temperature, is regarded as a linear function of past values and error terms. An ARIMA(p,d,q) model is a combination of AR(p), difference(d) and MA(q). The model for each environmental factor can be mathematically expressed as

$$X_t^i = a_t^i + \varphi_1^i X_{t-1}^i + \varphi_2^i X_{t-2}^i + \cdots + \varphi_p^i X_{t-p}^i - \varepsilon_t^i - \theta_1^i \varepsilon_{t-1}^i - \theta_2^i \varepsilon_{t-2}^i - \cdots - \theta_q^i \varepsilon_{t-q}^i \quad (1)$$

Where X_t^i is the i^{th} influence factor obtained by differencing d times, $\varphi_1^i,\ \varphi_2^i, \ldots \varphi_p^i$, $\theta_1^i, \theta_2^i, \ldots \theta_q^i$ are autoregressive and moving average coefficients need to be calculated for i^{th} factor. $\varepsilon_t^i,\ \varepsilon_{t-1}^i, \ldots \varepsilon_{t-q}^i$ are assumed to be independently and identically distributed with a mean of zero and a constant variance. Fig. 1 shows the flow diagram of ARIMA model, and the implementations can be described as follows:

Step 1: Check the stationarity of all the influence factors' historical data respectively since the ARIMA model is applicable only in stationary time series. A stationary time series has a constant mean as well as a constant variance. If the data is not stationary, a differencing operation is performed. If the data is still non-stationary, differencing is repeatedly performed until the data is finally in stationary. Augmented Dickey–Fuller (ADF) test [31] is used in this step. If ADF t-Statistics is lower than the Mackinnon critical value under the significant level of 5%, the series is supposed to be stationary.

Step 2: Estimate parameters for ARIMA(p,d,q), while "d" refers to the number of differencing derived by *Step 1*. As for (p,q), Auto-Correlation Function (ACF) and Partial Auto Correlation Function (PACF) graphs are plotted for qualitative analysis. If the ACF declines exponentially and the PACF spikes on the first p lags, the model can be autoregressive process like ARIMA(p,0,0). In contrast, for moving average processes ARIMA(0,0,q), the ACF spikes on the first q lags and the PACF declines exponentially. However, it is not accurate to estimate p and q only by ACF and PACF graphs. There are several criteria like Akaike Info Criterion (AIC) and Schwarz Criterion (SC) can be employed in quantitative analysis to obtain the specific values.

Step 3: Validate the ARIMA model obtained in *Step 2* and test on the residuals. If the residual term shows a normal distribution behavior with constant variance and zero mean, then it resembles white noise error and there is no need for further ARIMA modeling [28].

Step 4: Predict future data for each environmental factor.

3.2. Deep belief network

DBN is a type of deep neural networks [27], which can be viewed as a composition of simple, unsupervised networks such as restricted Boltzmann machines (RBMs). The structure of DBN is showed in Fig. 2. The connection of RBM's units is restricted to different layers, while no connection exists within the units in the same layer. A RBM involves an input layer (visible layer) and a hidden layer, while the hidden layer serves to detect the feature of the data from visible layer according to the connection weights W.

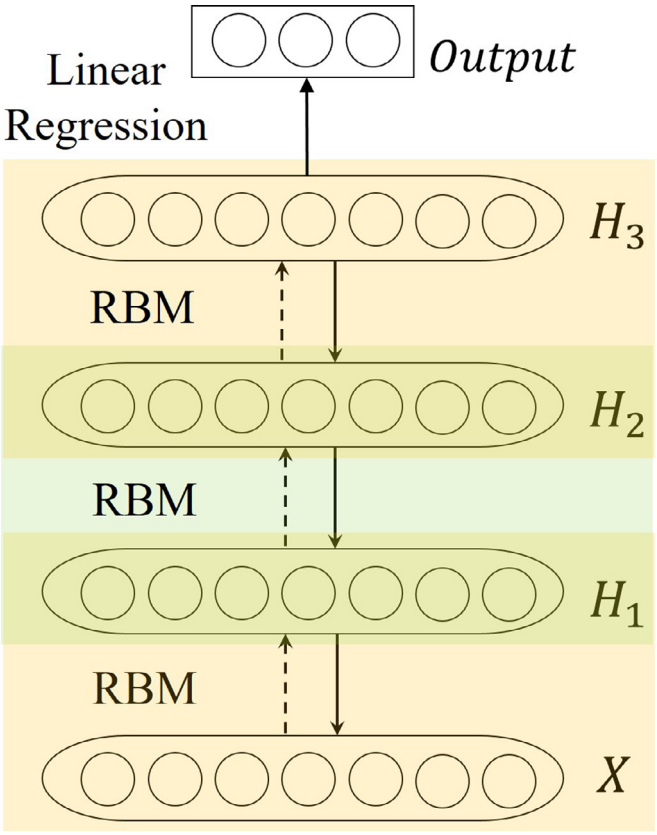

Fig. 2. The architecture of DBN. DBN consists of multiple RBMs, and the top layer is the linear regression.

This can be mathematically expressed as Eqs. (2) and (3) where v is the input vector and h is the hidden vector. The energy-based model RBM models the joint distribution between visible vector v and hidden vector h as Eq. (4) where $E(v, h)$ denotes the energy of a joint configuration of the hidden and visible units. After we get the joint distribution $P(h, v)$, we can calculate the marginal distribution $P(v)$, $P(h)$. Given a visible vector v, the hidden vector is a conditional probability $P(h|v)$. We analogously calculate the conditional probability $P(v|h)$.

$$P(h|v) = \frac{P(h, v)}{P(v)} \quad (2)$$

$$P(v|h) = \frac{P(v, h)}{P(h)} \quad (3)$$

$$P(h, v) = P(v, h) = \frac{1}{Z} e^{-E(v,h)} \quad (4)$$

$$Z = \sum_{v,h} e^{-E(v,h)} \quad (5)$$

Where h is the hidden units and v is the visible units.

Pre-training

The training process of RBMs is as follows [27]:

Step 1: Initialize weights W with normal distribution. For each input data x_t, $t \in [1, z]$, $v = x_t$.

Step 2: Compute the probability of hidden units $P(h|v)$.

$$P(h|v) = \sigma(Wv) \quad (6)$$

Step 3: Compute the probability of reconstructed visible units $P(v|h)$.

$$P(v|h) = \sigma(Wh) \quad (7)$$

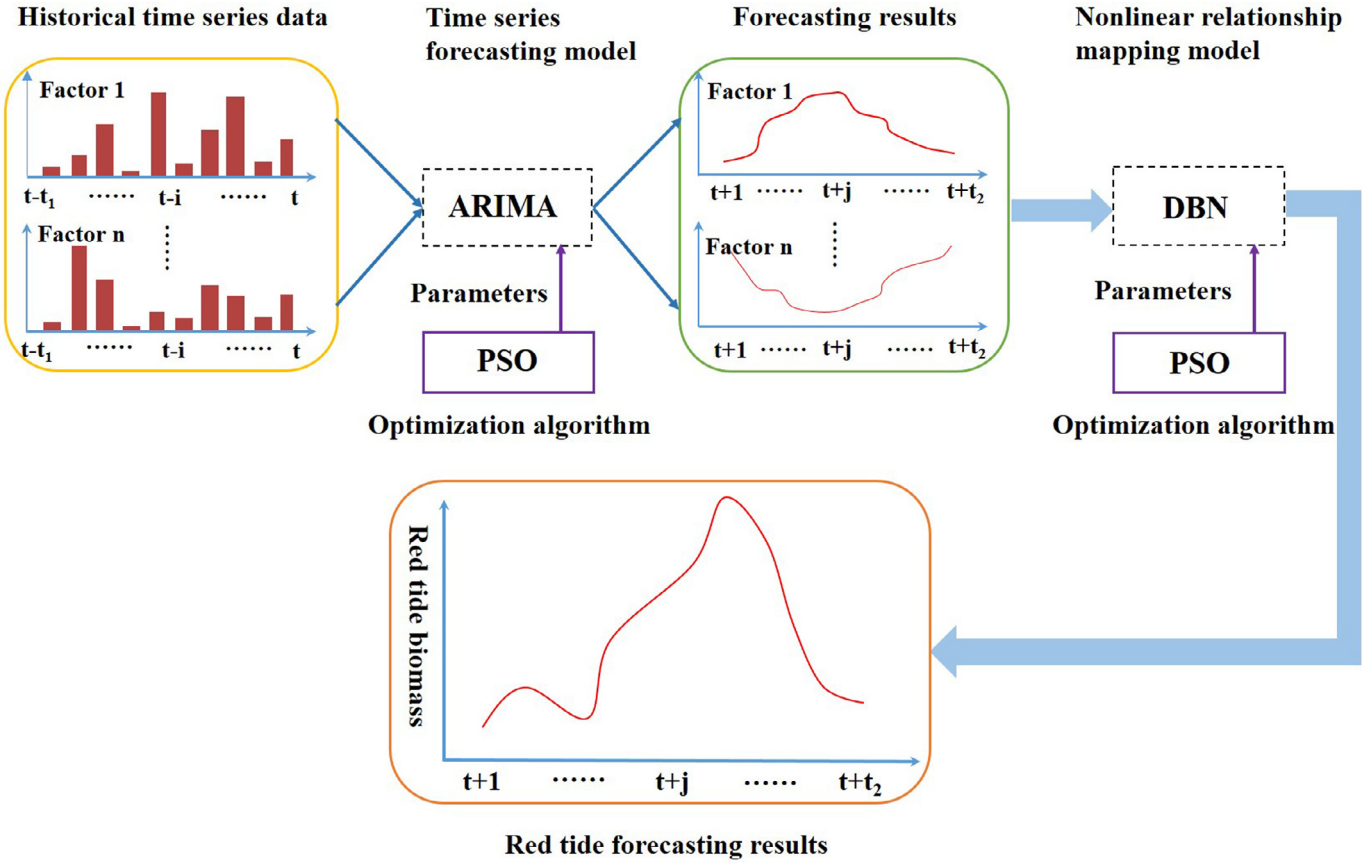

Fig. 3. The framework of ARIMA-DBN. ARIMA model is built for environmental factors to describe the temporal correlation and spatial heterogeneity of red tide. DBN serves to capture the complex nonlinear relationship between the environmental factors and the red tide biomass. PSO is used to seek for the optimum parameters for both ARIMA model and DBN.

Step 4: Obtain the reconstruction error ΔW.

$$\Delta W = \langle vh \rangle_{data} - \langle vh \rangle_{recon} \tag{8}$$

Step 5: Update the weights W, and calculate the energy function $E(v, h)$.

$$W \leftarrow W + \varepsilon(\Delta W) \tag{9}$$

$$E(v,h) = -\sum_i v_i b_i - \sum_j h_j c_j - \sum_{i,j} v_i h_j w_{ij} \tag{10}$$

b_i, c_j are the bias.

Step 6: Repeat *Step 2*~ *Step 5* until the energy function $E(v, h)$ decreases to be a convergent state.

Fine tune

Step 1: Train the first RBM with the input data x being the visible units.

Step 2: Fix the weights and bias of the first RBM, and the state of the hidden units is used as the input of the second RBM.

Step 3: The second RBM is stacked on the top of the first RBM after training.

Step 4: Iterate (2 and 3) for the desired number of layers, each time propagating upward either samples or mean values.

Step 5: Fine-tune all the parameters in this deep architecture with respect to a proxy for the DBN log-likelihood. In this work, the red tide biomass is continuous, so the fine-tune algorithm of DBN is a linear regression algorithm.

There are some advantages that distinguish DBN from other neural networks [29]:

1. have a higher modeling capacity per parameter.
2. have a fairly efficient training procedure that combines unsupervised generative learning for feature detection with a subsequent stage of supervised learning that fine tunes the features to optimize the discrimination.

3.3. A hybrid model for red tide forecasting

Red tide is a complicated nonlinear dynamic system influenced by several aspects, especially weather conditions, pollutant, and physical and chemical factors. The intricate relationships of all the environmental factors make it difficult to predict the occurrence of red tide. Nevertheless, a novel hybrid model ARIMA-DBN, as showed in Fig. 3, is proposed to deal with this hard problem. This model combines the strong forecasting ability of ARIMA and the powerful expression ability of DBN for nonlinear relationships.

For each influence factor, the historical development law is analyzed and then establishes the corresponding ARIMA model respectively due to the diversity among the factors. Subsequently, the future value of each factor is forecasted. In addition, the development law for the same factor is different in different areas. Therefore, we also need to change the parameters of the ARIMA model. The different ARIMA models reflect the temporal and spatial consistency of the environmental factors.

A DBN is trained to find the nonlinear relationship between all the environmental factors and the red tide biomass. After obtaining the prediction environmental factors by ARIMA, the future values are applied as the input to the DBN, and the output of DBN is the red tide biomass. Finally, the occurrence and the development of red tide can be estimated.

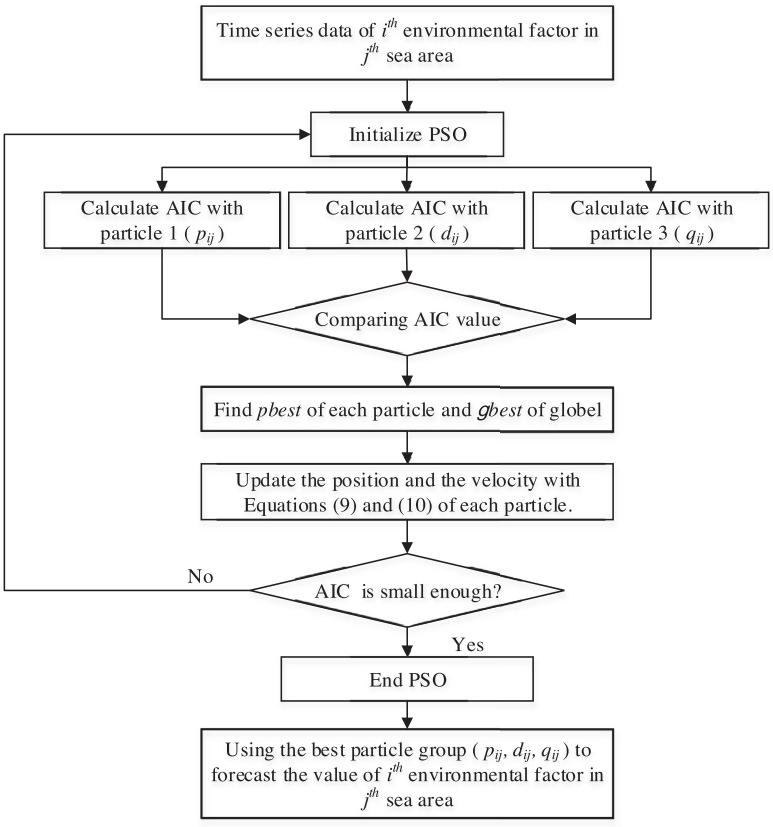

Fig. 4. Using the best particles to design ARIMA models.

3.4. Optimization algorithm

Particle swarm optimization (PSO) is a simple but effective stochastic optimization algorithm developed by Eberhart and Kennedy [30], inspired by social behavior of bird flocking. The process of PSO involves the following five steps:

Step 1: Initialize the particles (solutions) with start position $s^{k=0}$ and the start velocity $u^{k=0}$;

Step 2: Calculate the fitness with the chosen fitness function F of each particle;

Step 3: Find the best fitness each particle has achieved by far, and the best fitness is called *pbest*;

Step 4: Find another "best" value, which is a global best and called *gbest*; Update the position and the velocity with Eqs. (11) and (12) of each particle.

$$u_j^k = wu_j^{k-1} + c_1 r_1\left(pbest_i - s_i^{k-1}\right) + c_2 r_2\left(gbest_i - s_i^{k-1}\right) \tag{11}$$

$$s_i^k = s_i^{k-1} + u_j^{k-1} \tag{12}$$

Where w is the inertia factor, c_1 and c_2 are learning factor, r_1 and r_2 are stochastic function.

Step 5: If the algorithm does not reach the end condition (reach the maximum number of iterations K_{max} or the optimum fitness is smaller than the threshold), back to the *Step 2* and continue.

Since many parameters in the hybrid ARIMA-DBN model need to be decided, such as the (p, d, q) of each ARIMA model; the number of hidden layers, the units for each hidden layer of the DBN. Thus, PSO is applied to seeking the optimum solutions for time series forecasting and training the deep belief network. PSO can seek the optimum solutions without analyzing all the situations to enhance the training speed of the hybrid model.

Specifically, as Fig. 4 shows, the best parameter group (p_{ij}, d_{ij}, q_{ij}) can be obtained quickly through PSO algorithm. Here, $i = 1,2,\ldots,n$ (n is the number of selected factors), means the i^{th} influence factor, and $j = 1,2$, means the j^{th} sea area (Zhoushan & Wenzhou). While the AIC value is small enough, the parameter group is supposed to have the best performance in time series forecasting.

In terms of the DBN training, PSO is also adopted in order to find the optimum number of hidden layers H as well as the number of units m_l ($l = 1,2,\ldots H$) for each hidden layer. The flowchart is showed in Fig. 5 *MSE* is used here to decide whether the structure of DBN can predict red tide precisely.

4. Experiments

4.1. Study area and data pre-processing

4.1.1. Study area and data

Our experiment is based on the real dataset of Zhoushan coastal area (29°32′~30°57′N, 121°30′~123°23′E) and Wenzhou coastal areas (27°03′~28°21′N, 120°25′~121°50′), Zhejiang province, China. The Zhoushan and Wenzhou are sub-sea areas of Zhejiang coastal area, as Fig. 6 shows. The hydrodynamic environment of Zhejiang coastal area is complicated. Affected by the warm current and monsoon, the nutrient concentration in water is high. Coupled with water pollution, the nutrient content is increased, making the red tide biological reproduction more favorable. Furthermore, Zhoushan and Wenzhou coastal areas are sensitive areas of red tide in Zhejiang coastal area, and these two areas have substantial differences in spatial location and water environment. Therefore, we choose Zhoushan and Wenzhou as our case studies to evaluate the robustness and practicability of the proposed approach.

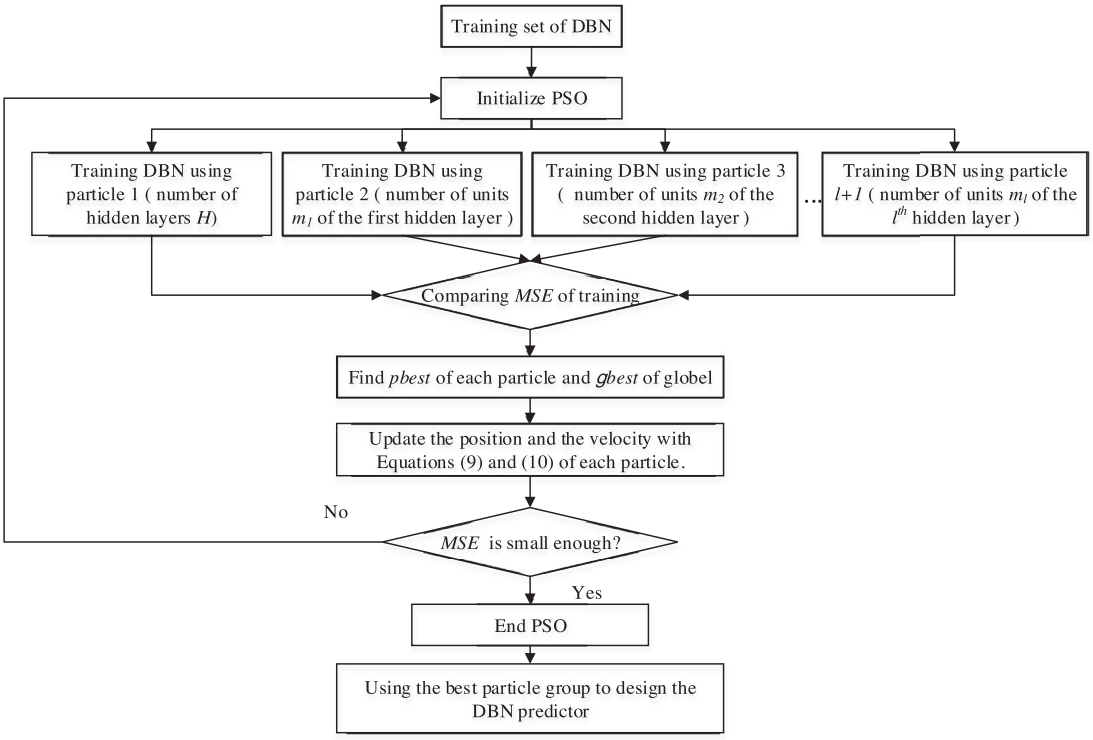

Fig. 5. Using the best particles to design an optimized predictor of DBN.

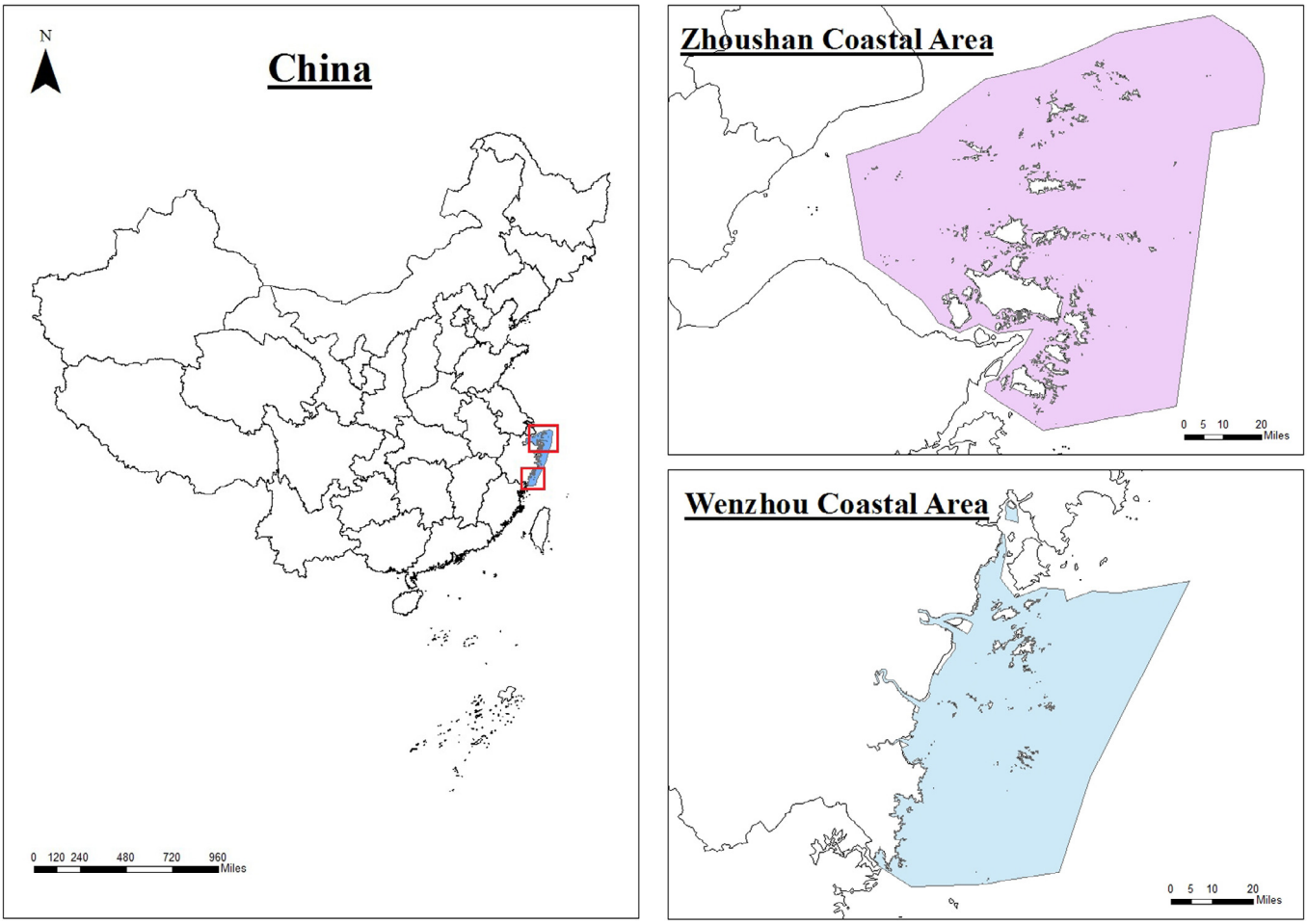

Fig. 6. Study area map. Zhoushan coastal area and Wenzhou coastal area are the study areas.

The data used in this paper come from the ship monitoring data during 2008–2014 in Zhoushan & Wenzhou coastal areas. Since lots of environmental factors have effect on red tide, the number, quality and the correlation degree with red tide of the selected factors used in the prediction model are of great importance to the final forecasting results. In this paper, we choose 12 factors including pH, dissolved oxygen, dissolved oxygen saturation, chemical oxygen demand (COD), temperature, chlorophyll, phosphate, ammonia, nitrate, nitrite, salinity, and silicate as the final influence factors. Each of these factors plays an indispensable role in the process of the growth, reproduction, extinction of the red tide algal. For example, ammonia, nitrate, nitrite and phosphate provide nutrition for the growth of red tide algal; pH, temperature, dissolved oxygen saturation and COD create good water environmental for red tide; chlorophyll can reflect the severity of the red tide.

A total of 6200 time series data (3800 in Zhoushan and 2400 in Wenzhou) is available in this work, and the data is collected weekly. Since the time interval is relatively large, the red tide biomass always steep rise/ drop without a middle record.

4.1.2. Data pre-processing

DBN is particularly good at extracting features and has strong expression ability of nonlinear relationships. However, we need to normalize the input and output data before training or using the DBN model. The reasons are listed as follows:

1) Eliminating the influence of range. The numeric range fluctuates widely, very small or very large for example, leading to slow convergence and long training time of the DBN.
2) Unifying the data range of factors, which can effectively avoid the influence of the input data range on output.
3) Unifying with the range of activation function.

The activation function of DBN is Sigmoid Function, and the range of the function is [0,1]. Thus, the output (red tide biomass) need to be normalized to the same range of the function. Owing to the large-range of the red tide biomass in the real scale, which is from 10^2 to 10^7, the common normalization method is invalid. Therefore, we propose a new normalization Eq. (13), which can credibly retain the characteristics of the original data. As for the environmental factors, they are normalized as Eq. (14).

$$y^* = \frac{\log(y)}{\max(\log(y))} \tag{13}$$

$$y^* = \frac{y}{\max(y)} \tag{14}$$

Where y^* is the normalized value of y.

4.2. Evaluation criteria

The fitting degree (R), root mean squared error (*RMSE*), the mean absolute error (*MAE*) and the mean absolute percentage error (*MAPE*) are used to evaluate the predictive ability of the hybrid model.

$$\begin{cases} R = 1 - \sqrt{\frac{\sum_{i=1}^{N}(y_i - y'_i)^2}{\sum_{i=1}^{N}(y_i)^2}} \\ RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - y'_i)^2}{N}} \\ MAE = \frac{\sum_{i}^{N}|y_i - y'_i|}{N} \\ MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - y'_i}{y_i}\right| \times 100\% \end{cases} \tag{15}$$

Where y_i is the actual value and y'_i is the predicted value.

The important properties of the above four evaluation criteria are as follows:

Table 1
Parameters of different environmental factors for ARIMA.

ENVIRONMENTAL FACTORS	Zhoushan			Wenzhou		
	p	d	q	p	d	q
pH	**0**	**0**	**1**	**4**	**0**	**6**
TEMPERATURE	6	0	4	2	0	5
SALINITY	2	0	6	6	0	4
COD	**4**	**1**	**4**	**4**	**0**	**5**
DISSOLVED OXYGEN	6	0	6	4	0	6
DISSOLVED OXYGEN SATURATION	**5**	**0**	**3**	**0**	**0**	**3**
PHOSPHATE	**2**	**0**	**2**	**6**	**1**	**6**
NITRITE	**4**	**0**	**6**	**2**	**1**	**1**
NITRATE	5	0	6	4	0	3
AMMONIA	5	0	2	1	0	0
SILICATE	5	0	5	2	0	6
CHLOROPHYLL	**0**	**0**	**1**	**2**	**0**	**3**

R measures the fitting degree between the actual series and the prediction series. The range of R is [0,1], and the closer to 1, the better effect of the forecasting model.

RMSE is the square root of calculated mean squared error (*MSE*), which measures the average squared deviation of forecasted values and gives an overall idea of the error occurred during forecasting. RMSE is sensitive to the changes of scale and data transformations as well.

MAE measures the average absolute deviation of forecasted values from original ones, and it shows the magnitude of overall error where the effects of positive and negative errors are not eliminated. For a good forecast, the obtained MAE should be as small as possible.

MAPE represents the percentage of average absolute error. It is independent of the scale of measurement, but affected by data transformation.

RMSE, *MAE* and *MAPE* are commonly used to measure the performance of model. In this work, they are used to evaluate the forecasting accuracy of the hybrid model ARIMA-DBN. The more close to 0, the higher accuracy the algorithm is. Besides, the four criteria can be applied to compare different models.

4.3. Environmental factors forecasting

4.3.1. ARIMA parameters determination

There are 12 factors in Zhoushan and Wenzhou need to be analyzed respectively. Since the data interval is one week, we use the previous 20 frames as historical data, and then applied to the multi-step forecasting of the next 10 frames of the corresponding values, which means nearly 5 months historical data are analyzed to forecast more than 2 months data ahead in real records.

PSO algorithm is applied to seek the optimum parameter groups of ARIMA models. As Fig. 4 shows, particle 1 refers to the parameter p of AR; particle 2 refers to the differing times d; and particle 3 refers to the parameter q of MA. The range of the parameters are restricted to $p \in \{0, 1, \ldots, 10\}$, $q \in \{0, 1, \ldots, 10\}$, $d \in \{0, 1, 2\}$. PSO algorithm is run 24 (12 environmental factors $\times$ 2 coastal areas) times in total to obtain all the parameters. Finally, the results are as follows.

From the table above, we can see that the parameters of the same environmental factor in different areas are of great diversity. The models of pH and chlorophyll in Zhoushan area are MAs, while in Wenzhou are ARMAs. For dissolved oxygen saturation, the MA model is applicable in Wenzhou, but in Zhoushan, the model ought to be ARMA. Besides, for phosphate and nitrate, the time series in Zhoushan are stationary while in Wenzhou are not, and as for COD, the situation is quite the contrary.

Table 1 indicates that the development law of the same factor is diverse from areas, demonstrating obvious spatial heterogeneity.

Table 2
The time series forecasting performance of each environmental factor.

ENVIRONMENTAL FACTORS	Zhoushan				Wenzhou			
	R	*MAE*	*RMSE*	*MAPE*	*R*	*MAE*	*RMSE*	*MAPE*
pH	0.991	0.052	0.067	0.63%	0.990	0.052	0.076	0.63%
TEMPERATURE	0.922	1.472	1.718	6.69%	0.818	3.584	4.564	12.91%
SALINITY	0.977	0.450	0.596	1.64%	0.905	1.883	2.333	7.59%
COD	0.608	0.127	0.159	29.20%	0.541	0.269	0.407	32.82%
DISSOLVED OXYGEN	0.916	0.454	0.577	6.49%	0.886	0.654	0.758	9.44%
DISSOLVED OXYGEN SATURATION	0.917	5.931	7.878	6.38%	0.934	4.573	6.205	4.82%
PHOSPHATE	0.822	0.003	0.004	13.08%	0.580	0.009	0.011	62.84%
NITRITE	0.757	0.002	0.003	18.67%	0.437	0.008	0.011	45.39%
NITRATE	0.682	0.154	0.179	25.77%	0.764	0.087	0.110	21.12%
AMMONIA	0.785	0.003	0.005	12.34%	0.736	0.006	0.010	14.05%
SILICATE	0.907	0.049	0.074	5.49%	0.629	0.253	0.350	41.37%
CHLOROPHYLL	0.458	2.422	3.756	29.37%	0.222	0.889	1.327	68.89%

Therefore, it is of great necessity to establish different ARIMA forecasting models for different coastal areas.

4.3.2. Forecasting results of environmental factors

The time series forecasting models determined in Table 1 are applied to predict the future trend for all the factors. For each coastal area, 6 groups of experiments are executed, and the data samples used for different groups are collected in different monitoring stations. The forecasting errors and the fitting degrees of the forecasting results are shown in Table 2. Since the space is limited, we just show the results of two groups of each area in Figs 7 and 8.

Table 2 presents the results of time series forecasting of each factor in Zhoushan coastal area and Wenzhou coastal area. As a whole, the future values of many factors are well forecasted, such as the *R* of pH, dissolved oxygen saturation and salinity are higher than 0.9. Nevertheless, there are some factors with very low *R* and high errors. For example, the forecasting *R* of chlorophyll, COD in both areas are below 0.7, which may have impact on the future red tide forecasting results.

While comparing the two areas, forecasting results of Zhoushan are better than Wenzhou because there are six factors with the *R* lager than 0.9 in Zhoushan while there are only three in Wenzhou. Except dissolved oxygen saturation and nitrate, the prediction errors in Wenzhou are all larger than Zhoushan.

4.4. Current prediction of red tide with DBN

As Fig. 3 shows, PSO is applied to find the optimum structure of DBN. The detailed steps are shown in Fig. 5. The particle 1 is the number of hidden layers *n*, the rest particles are the number of units $m_l (l = 1,2,\ldots,H)$ of hidden layers. If *n* is large, DBN is more likely to be over fitting, which results in inaccurate prediction. Thus, according to the complexity of the problem, the range of *n* is restricted as $H \in [1, 5]$. Besides, according to the empirical equation, $m_l \in [3, 14]$.

The position and the velocity of the particles in PSO algorithm are updated with the criterion of MSE, and it can reach the optimum solutions without calculating all the situations.

All the 6200 samples are randomly divided into 3:1:1(training samples: validation samples: testing samples). The optimum structure decided by PSO is $H=2$, $m_1 = 8$, $m_2 = 6$, and the testing results is shown in Table 3. Furthermore, 300 samples are randomly selected in each area, the actual value and predicted value are shown in Figs. 9 and 10.

As we can see from the table, DBN can well predict the red tide biomass from the input factors with the average *R* of the two areas reaching 0.925, and the three kinds of errors are small. From the Figs. 9 and 10, we can find that the peak values of red tide biomass are precisely predicted, which is important in red tide earn-warning. When the red tide biomass higher than the threshold value, red tide occurs.

Table 3
Results of red tide prediction with DBN.

Sea area	*R*	*MAE*	*RMSE*	*MAPE*
Zhoushan	0.948	0.028	0.037	4.00%
Wenzhou	0.901	0.047	0.059	8.74%
Average	0.925	0.038	0.048	6.37%

Table 4
The red tide biomass forecasting performance in different areas.

Sea area	Experiments	*R*	*MAE*	*RMSE*	*MAPE*
Zhoushan	1	0.826	0.130	0.144	15.41%
	2	0.802	0.138	0.163	16.53%
	3	0.855	0.110	0.118	13.51%
	4	0.839	0.098	0.132	11.53%
	5	0.851	0.108	0.121	13.28%
	6	0.830	0.112	0.138	14.29%
	Average	0.832	0.117	0.137	14.09%
Wenzhou	1	0.690	0.198	0.238	27.14%
	2	0.829	0.091	0.121	14.39%
	3	0.771	0.123	0.158	20.54%
	4	0.783	0.109	0.149	17.02%
	5	0.770	0.119	0.148	21.53%
	6	0.754	0.138	0.176	21.32%
	Average	0.759	0.129	0.169	20.32%
Total	Average	0.798	0.123	0.154	17.20%

4.5. Time series forecasting of red tide

4.5.1. ARIMA-DBN hybrid model

The forecasted values of the 12 factors obtained from Section 4.3 are used as input, and the corresponding red tide biomass values are predicted by DBN model which is obtained in Section 4.4. The forecasting results are shown in Figs. 11 and 12.

From the two figures (Figs. 11 and 12), the model can well forecast the red tide biomass. Besides, the model is very sensitive to peak value, with the forecasted peak even higher than the actual, which is of great significance to the red tide early warning. The detailed error and fitting degree between predicted data and actual data for each experiment group are evaluated as Table 4 shows.

In total, the *R* between the predicted value and actual value is 0.798, and all the results of the three error evaluation criteria are very small, with the *MAE* equaling 0.123, *RMSE* equaling 0.154 and *MAPE* equaling 17.2%. It is proved that the proposed hybrid model ARIMA-DBN, which consists of ARIMA model and Deep Belief Network, can reliably forecast the future trend of red tide biomass through historical data.

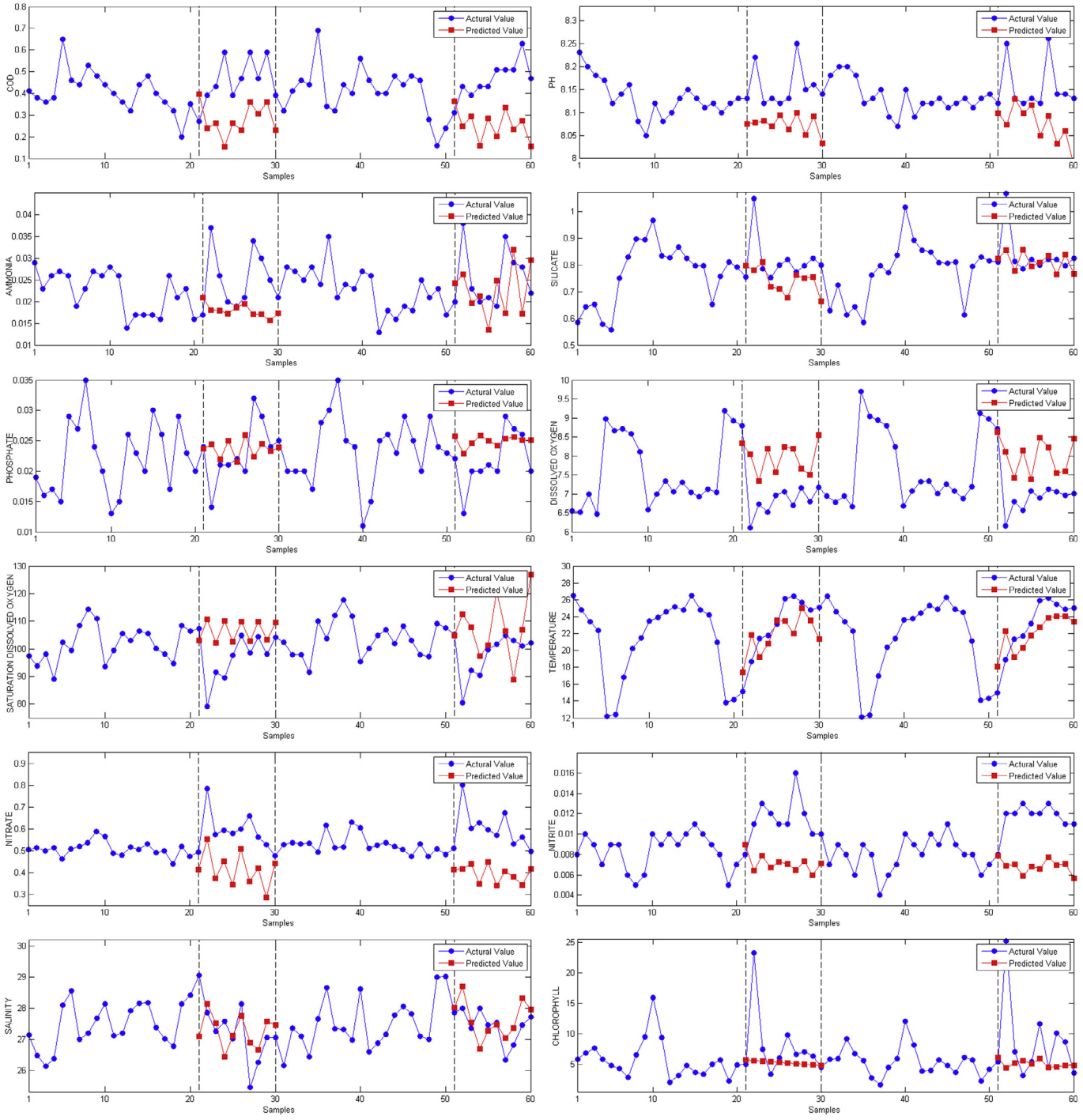

Fig. 7. The time series forecasting results of 12 environmental factors in Zhoushan area. The 20 frames historical data is analyzed by ARIMA model, and the future 10 frames data is forecasted.

Comparing the two areas, it is found that the forecasting results of Zhoushan is better than Wenzhou, as the mean R of six groups in Zhoushan is 0.832, much higher than 0.759 in Wenzhou. The mean errors are smaller in Zhoushan than in Wenzhou. Hence, the spatial heterogeneity of red tide is confidently demonstrated.

In summary, when establishing the time series forecasting model, taking temporal and spatial heterogeneity into consideration is of great necessity. First of all, it can help improve the accuracy of the time series forecasting on each factor. Then, the learning burden of the DBN can be reduced, since DBN only need to consider the nonlinear relationship between 12 influence factors and red tide biomass. As a result, the accuracy and reliability of the DBN can be improved and it is less likely to be over fitting.

4.5.2. Results of different consecutive frames

For different future forecasting consecutive frames, the prediction performance is changeable, as shown in Fig. 13. With the increase of consecutive frames, the R of different areas are all increased first and then decreased. When the consecutive frame is 5, the R reaches to the maximum. The MAE, RMSE and MAPE of all areas are in the opposite. With the increase of consecutive frames, the three kinds of prediction errors are all decreased first and then

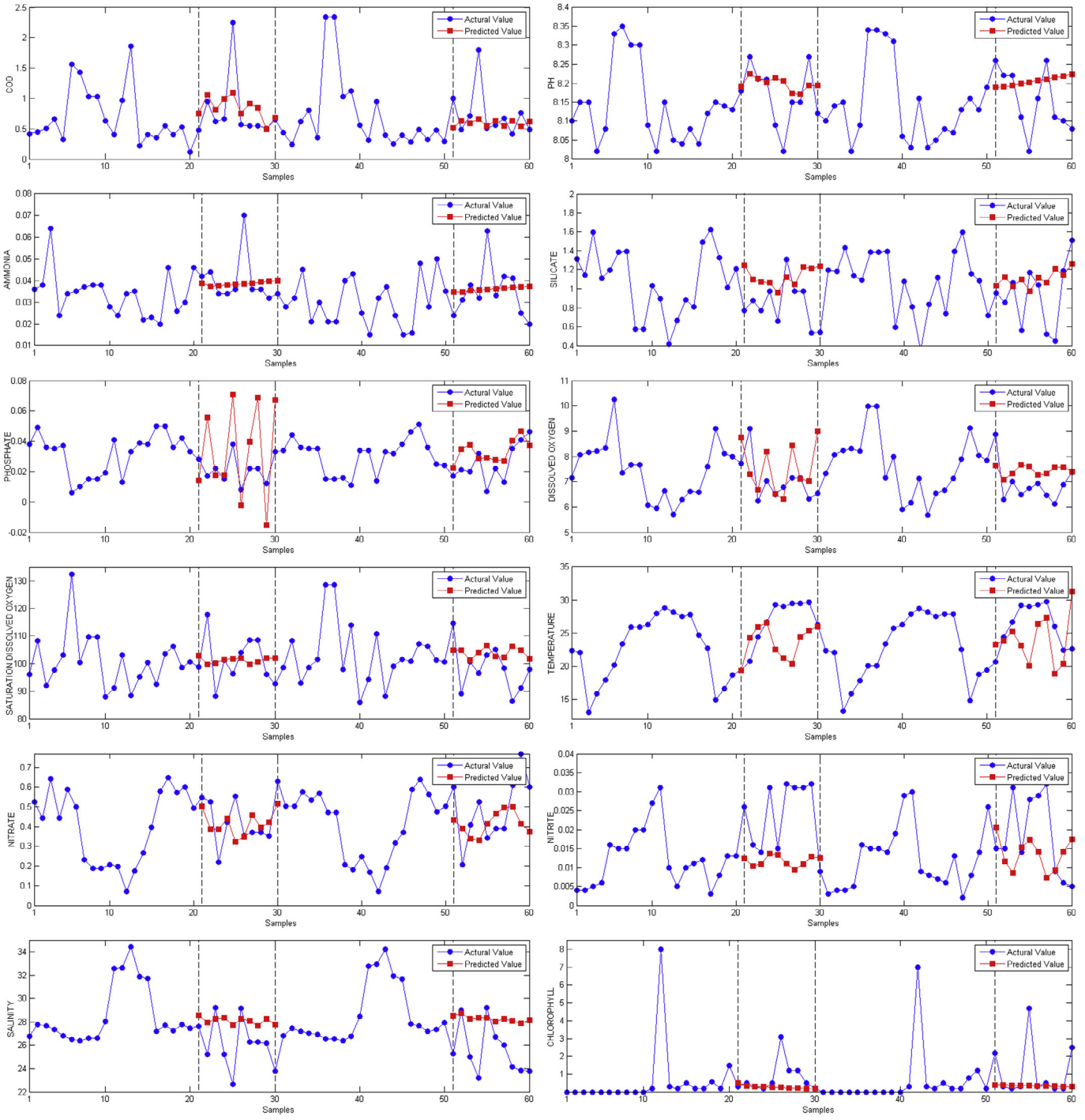

Fig. 8. The time series forecasting results of 12 environmental factors in Wenzhou area. The 20 frames historical data is analyzed by ARIMA model, and the future 10 frames data is forecasted.

increased. When the consecutive frame is 5, the three kinds of errors reaches the minimum. Therefore, when the consecutive frame is 5, the proposed model of ARIMA-DBN can forecast the red tide with great accuracy in the study area.

4.6. Algorithm comparison

For comparison, several algorithms are employed as comparison methods. The algorithm comparison part is divided into two groups. The first group is to compare the performance of current prediction of the red tide, and the second group is to compare the performance of time series forecasting of the red tide. The first group does not contain the time series forecasting, but only models the relationship between environmental factors and red tide. For each sample of the first group, the input dimension is 12, and the output dimension is 1. As for the second group, the input dimension of each sample is 240 (20 frames historical data of 12 environmental factors), and the output dimension is 10 (10 frames red tide of future), which predicts red tide using the historical data directly.

BP neural network [35], RBF neural network and GRNN are selected to compare with DBN in the first group. BP neural network

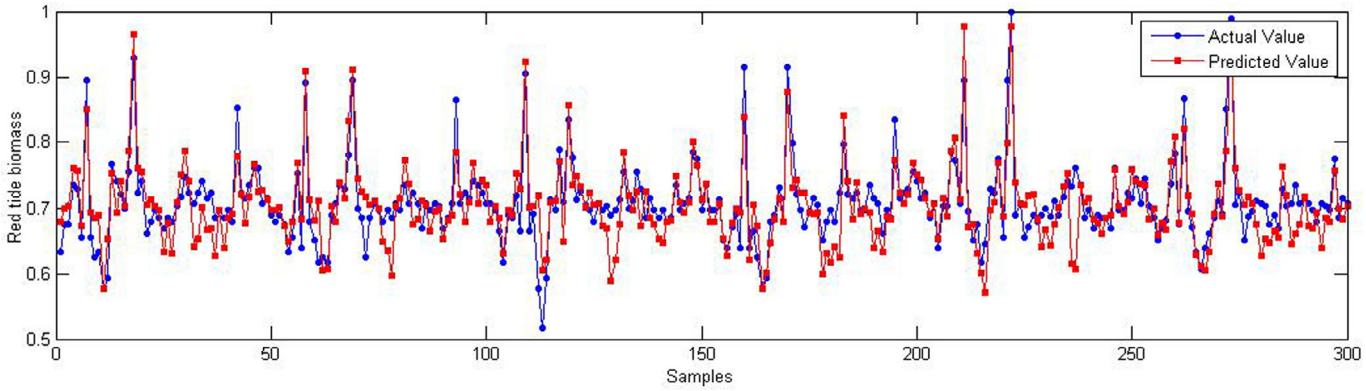

Fig. 9. Red tide biomass prediction results with DBN of Zhoushan area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

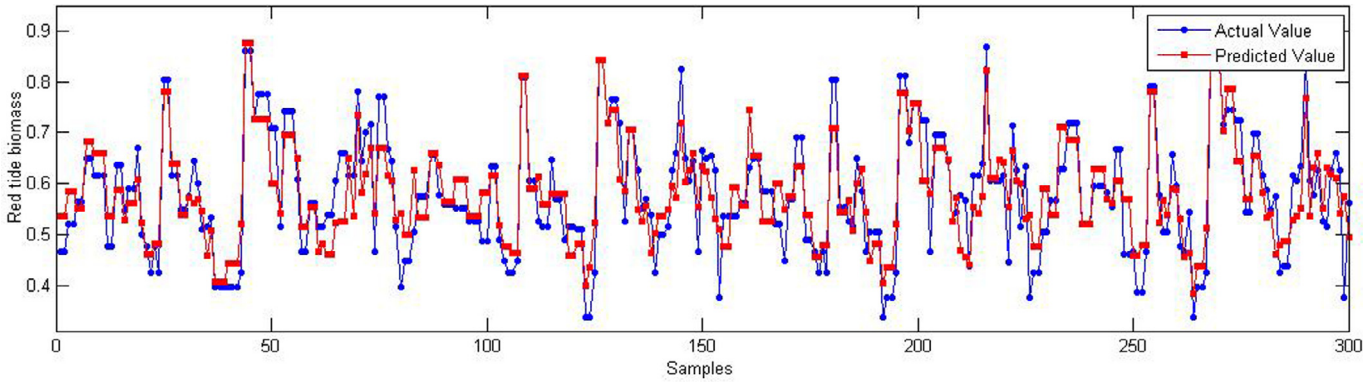

Fig. 10. Red tide biomass prediction results with DBN of Wenzhou area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

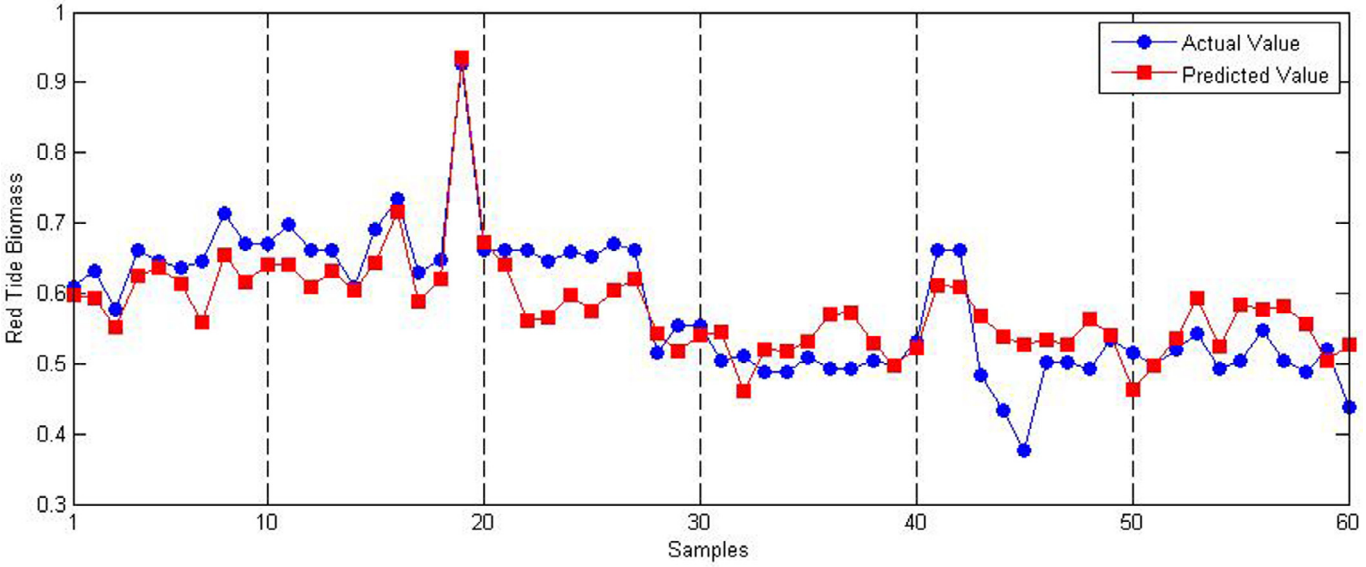

Fig. 11. Red tide biomass time series forecasting results of Zhoushan area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

is widely used. It plays an important role in various fields such as financial prediction, pattern recognition, and classification. RBF neural network is a kind of feedforward network. It provides an effective means for data prediction and modeling with low computation cost and high learning speed. GRNN is a normalized RBF network, and can also be considered as a one-pass learning algorithm, with a highly parallel structure. GRNN has been widely used in signal process, structure analysis, and decision analysis [9].

The comparison results is shown in Table 5.

We can see that the results of DBN is much better than the other three methods. The R of DBN reaches 0.925 while the BP neural network is 0.901, and the R of the rest two methods are both lower than 0.9. All the errors of DBN are smaller than others.

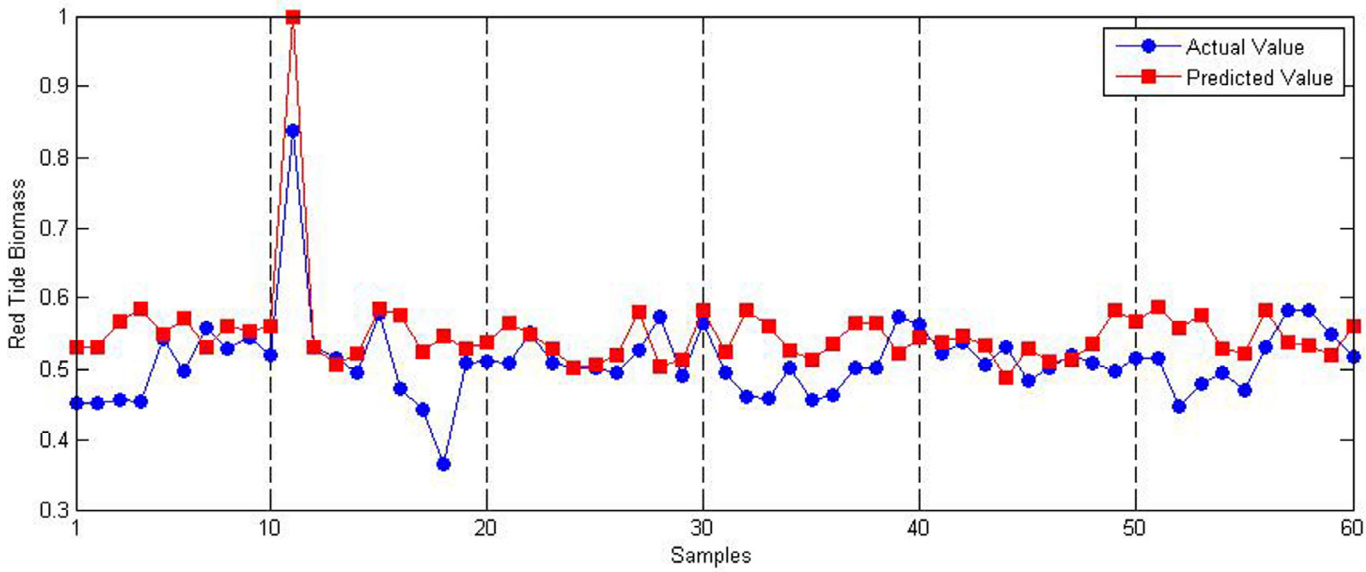

Fig. 12. Red tide biomass time series forecasting results of Wenzhou area. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

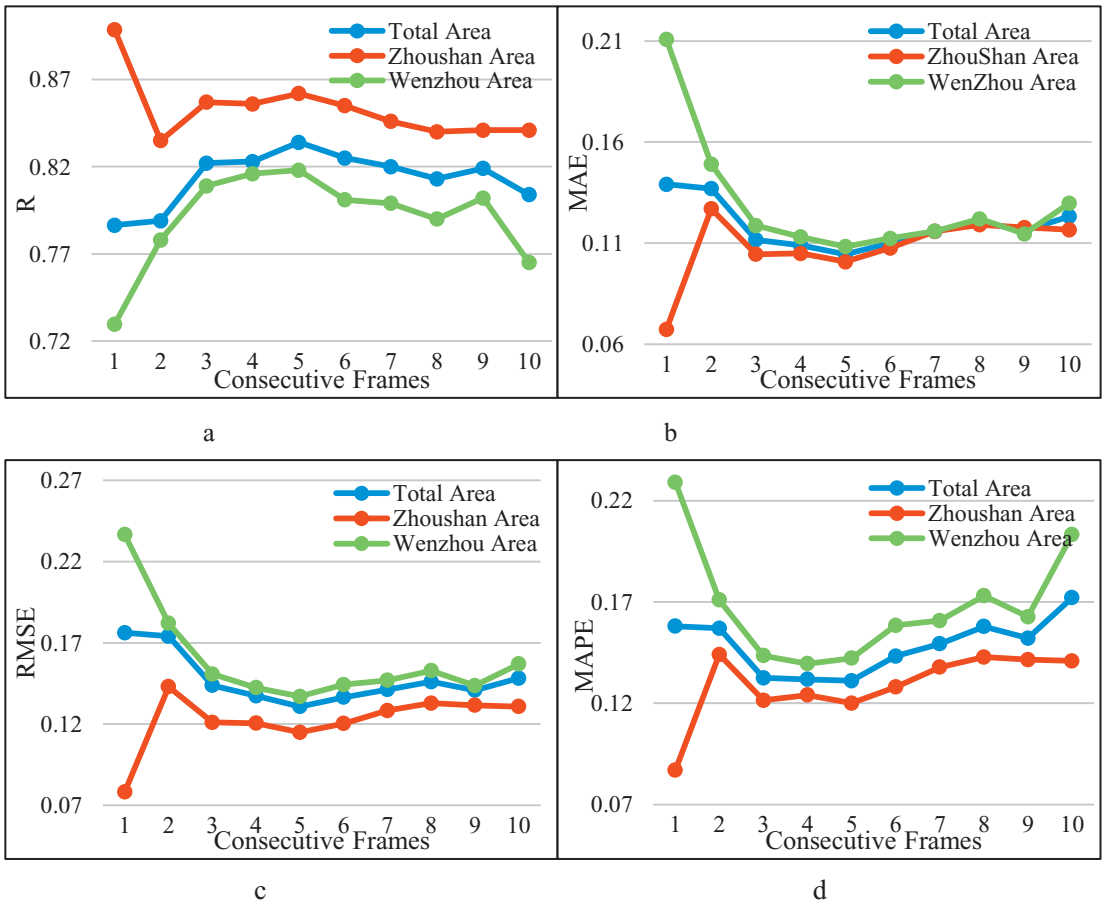

Fig. 13. The forecasting performance of different consecutive frames. a is the comparison of R, b,c,d are the comparison of different error.

The results demonstrate that the proposed method DBN achieves the best on describing the complex nonlinear relationship between the red tide and the environmental factors.

In order to verify the significance of the proposed ARIMA-DBN model, DBN, BP neural network (In order to distinguish with the models used in the first group, these two models are written as DBN* and BP* below) and typical hybrid model ARIMA-BP [10] are employed as comparison approaches of the second group. The results are shown in Table 6.

It is obvious that the forecasting performance of ARIMA-DBN is better than other approaches. The *R* of ARIMA-DBN is 0.798, much higher than ARIMA-BP, DBN* and BP*, stating that the hybrid

Table 5
Red tide current prediction algorithms comparison.

Algorithm	*R*	*MAE*	*RMSE*	*MAPE*
BP	0.901	0.046	0.065	8.21%
RBF	0.858	0.067	0.085	11.83%
GRNN	0.873	0.054	0.076	9.57%
DBN	**0.925**	**0.037**	**0.048**	**6.37%**

Table 6
Red tide time series forecasting models comparison.

Algorithm	*R*	*MAE*	*RMSE*	*MAPE*
ARIMA-DBN	**0.798**	**0.123**	**0.154**	**17.21%**
ARIMA-BP	0.716	0.164	0.192	28.30%
DBN*	0.654	0.198	0.217	36.47%
BP*	0.598	0.224	0.253	41.12%

model has good practicability on the real dataset. In addition, because DBN* has no capacity for learning long-term dependencies, it is not so good at forecasting time series, the *R* is 0.654, much lower than the current prediction result.

5. Conclusion

In this paper, a hybrid model of ARIMA-DBN is proposed to predict the red tide before occurrence. The ARIMA models are built for each environmental factor in different areas to describe the temporal correlation and spatial heterogeneity. Meanwhile, the DBN is used to seek the complex nonlinear relationship between environmental factors and red tide biomass. PSO is employed as the optimization algorithm to enhance the speed of seeking for the optimum parameters of ARIMA-DBN. Based on all above, the early warning of red tide can be achieved. The ship monitoring data during 2008–2014 of Zhoushan and Wenzhou coastal areas is used as the experimental dataset. The proposed ARIMA-DBN model is applied to forecast red tide in both Zhoushan and Wenzhou coastal areas. For different area like Zhoushan and Wenzhou, the model can well observe the spatial heterogeneity. As far as the experimental results can show, the proposed method works well in red tide forecasting, and has good practicability and robustness.

Acknowledgments

This research was supported by Public Science and Technology Research Funds' Projects (201305012, 201505003); the Fundamental Research Funds for the Central Universities from Ministry of Education of the People's Republic of China under grant number 2016XZZX004-02.

References

[1] S. Park, S.R. Lee, Red tides prediction system using fuzzy reasoning and the ensemble method, Appl. Intell. 40 (2) (2014) 244–255.
[2] X.M. Hu, D. Wang, H.W. Qu, X.R. Shi, Prediction Research of Red Tide Based on Improved FCM, in: Math. Probl. Eng., 2016, pp. 1–8.
[3] K. Davidson, D.M. Anderson, et al., Forecasting the risk of harmful algal blooms, Harmful Algae 53 (2016) 1–7.
[4] S.M. Gu, X.H. Sun, Y.H. Wu, et al., An approach to forecast red tide using generalized regression neural network, in: 2012 Eighth International Conference on Natural Computation (ICNC), 2012, pp. 194–198.
[5] G.E.P. Box, G.M. Jenkins, G.C. Reinsel, et al., Time Series Analysis: Forecasting and Control, Wiley, New York, 2015.
[6] Y.S. Lee, L.I. Tong, Forecasting time series using a methodology based on autoregressive integrated moving average and genetic programming, Knowl. Based Syst. 24 (1) (2011) 66–72.
[7] C.Y. Lam, W.H. Ip, C.W. Lau, A business process activity model and performance measurement using a time series ARIMA intervention analysis, Expert Syst. Appl. 36 (3) (2009) 6986–6994.
[8] M. Lippi, M. Bertini, P. Frasconi, Short-term traffic flow forecasting: an experimental comparison of time-series analysis and supervised learning, IEEE Trans. Intel. Transport. Syst. 14 (2) (2013) 871–882.
[9] B. Zhu, Y. Wei, Carbon price forecasting with a novel hybrid ARIMA and least squares support vector machines methodology, Omega 41 (3) (2013) 517–524.
[10] C.N Babu, B.E. Reddy, A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data, Appl. Soft Comput. 23 (2014) 27–38.
[11] C. Ren, N. An, J. Wang, et al., Optimal parameters selection for BP neural network based on particle swarm optimization: a case study of wind speed forecasting, Knowl. Based Syst. 56 (2014) 226–239.
[12] T. Zhou, S. Gao, J. Wang, et al., Financial time series prediction using a dendritic neuron model, Knowl. Based Syst. 105 (2016) 214–224.
[13] N.I. Sapankevych, R. Sankar, Time series prediction using support vector machines: a survey, IEEE Comput. Intel. Mag. 4 (2) (2009) 24–38.
[14] K. McAlinn, M. West, Dynamic bayesian predictive synthesis in time series forecasting. arXiv preprint: 1601.07463, (2016).
[15] J.F.L. Oliveira, T.B. Ludermir, A hybrid evolutionary decomposition system for time series forecasting, Neurocomputing 180 (2016) 27–34.
[16] I. Aizenberg, L. Sheremetov, L. Villa-Vargas, Multilayer neural network with multi-valued neurons in time series forecasting of oil production, Neurocomputing 8495 (2015) 61–70.
[17] F.M. Tseng, H.C. Yu, G.H. Tzeng, Combining neural network model with seasonal time series ARIMA model, Technol. Forecast. Soc. Change 69 (1) (2001) 71–87.
[18] L. Wang, H. Zou, J. Su, et al., An ARIMA-ANN hybrid model for time series forecasting, Syst. Res. Behav. Sci. 30 (3) (2013) 244–259.
[19] U. Yolcu, E. Egrioglu, C.H. Aladag, A new linear & nonlinear artificial neural network model for time series forecasting, Decis. Supp. Syst. 54 (3) (2013) 1340–1347.
[20] Q. Cai, D. Zhang, W. Zheng, et al., A new fuzzy time series forecasting model combined with ant colony optimization and auto-regression, Knowl. Based Syst. 74 (2015) 61–68.
[21] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, 2012.
[22] S. Yi, X.G. Wang, X.O. Tang, Deep learning face representation from predicting 10,000 classes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[23] C. Szegedy, W. Liu, Y. Jia, et al., Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[24] R. Collobert, J. Weston, A unified architecture for natural language processing: deep neural networks with multitask learning, in: Proceedings of the 25th International Conference on Machine Learning, ACM, 2008.
[25] A. Bordes, X. Glorot, J. Weston, et al., Joint learning of words and meaning representations for open-text semantic parsing, AISTATS 22 (2012) 127–135.
[26] K. Cho, B. Van. Merriënboer, C. Gulcehre et al, Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014).
[27] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554.
[28] J. Hassan, ARIMA and regression models for prediction of daily and monthly clearness index, Renew. Energy 68 (2014) 421–427.
[29] A. Mohamed, G.E. Hinton, G. Penn, Understanding how deep belief networks perform acoustic modelling, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4273–4276.
[30] J. Kennedy, in: Particle Swarm optimization, Encyclopedia of Machine Learning, Springer, US, 2011, pp. 760–766.
[31] R.S. Tsay, Analysis of Financial Time Series, Wiley, New York, 2005.
[32] K.T.M. Wong, J.H.W. Lee, P.J. Harrison, Forecasting of environmental risk maps of coastal algal blooms, Harmful Algae 8 (3) (2009) 407–420.
[33] C.K. Lee, T.G. Park, Y.T. Park, et al., Monitoring and trends in harmful algal blooms and red tides in Korean coastal waters, with emphasis on Cochlodinium polykrikoides, Harmful Algae 30 (2013) S3–S14.
[34] S. Park, S.R. Lee, J.H. Park, et al., Prediction of red tide blooms using decision tree model, in: IEEE International Conference on Convergence, 2011, pp. 710–713.
[35] L. Velo-Suárez, J.C. Gutiérrez-Estrada, Artificial neural network approaches to one-step weekly prediction of Dinophysis acuminata blooms in Huelva (Western Andalucía, Spain), Harmful Algae 6 (3) (2007) 361–371.
[36] H. Cui, J. Feng, J. Guo, et al., A novel single multiplicative neuron model trained by an improved glowworm swarm optimization algorithm for time series prediction, Knowl. Based Syst. 88 (2015) 195–209.
[37] T. Kuremoto, S. Kimura, K. Kobayashi, et al., Time series forecasting using a deep belief network with restricted Boltzmann machines, Neurocomputing 137 (2014) 47–56.
[38] M. Längkvist, L. Karlsson, A. Loutfi, A review of unsupervised feature learning and deep learning for time-series modeling, Pattern Recognit. Lett. 42 (2014) 11–24.
[39] X. Qiu, L. Zhang, Y. Ren, et al., Ensemble deep learning for regression and time series forecasting, in: IEEE Symposium on In Computational Intelligence in Ensemble Learning, 2014, pp. 1–6.
[40] M.L. Wells, V.L. Trainer, T.J. Smayda, et al., Harmful algal blooms and climate change: Learning from the past and present to forecast the future, Harmful Algae 49 (2015) 68–93.