Ecological Modelling 364 (2017) 42–52

Contents lists available at ScienceDirect

Ecological Modelling

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

An optimization of artificial neural network model for predicting chlorophyll dynamics

Wenchong Tian [a], Zhenliang Liao [a,*], Jin Zhang [b,*]

[a] Key Laboratory of Yangtze River Water Environment, Ministry of Education, College of Environmental Science and Engineering, Tongji University, 200092 Shanghai, China
[b] Institute of Urban Water Management, Technische Universität Dresden, 01062 Dresden, Germany

ARTICLE INFO

Article history:
Received 3 February 2017
Received in revised form 16 August 2017
Accepted 12 September 2017

Keywords:
Artificial neural networks
Chlorophyll dynamics
Algal bloom
Non-stationary time series

ABSTRACT

As one of the factors to represent some species of algae, chlorophyll dynamics model has been regarded as one of the early-warning proactive approaches to prevent or mitigate the occurrence of some algal blooms. To decrease the cost of aquatic environmental in-situ monitoring and increase the accuracy of bloom forecasting, a traditional artificial neural network (ANN) based chlorophyll dynamics prediction model had been optimized. This optimization approach was conducted by presenting the change of chlorophyll value rather than the base value of chlorophyll as the output variable of the network. Both of the optimized and traditional networks had been applied to a case study. The results of model performance indices show that the optimized network predicts better than the traditional network. Furthermore, the non-stationary time series was employed to explain this phenomenon from a theoretical aspect. The proposed approach for chlorophyll dynamics ANN model optimization could assist the essential proactive strategy for algal bloom control.

1. Introduction

It has been widely reported that lakes and reservoirs are commonly susceptible to eutrophication, and the subsequent algal blooms result in an adverse effect on the drink water security (Gu et al., 2017). Due to the adverse effect on the water quality, the eutrophication induced algal blooms can disrupt water supply to the surrounding cities (Zhang et al., 2014). Traditionally, algal blooms in-situ management programs require routine monitoring and/or reactive monitoring whereby blooms are monitored episodically and at a greater frequency only when a problematic algal species are detected, or a bloom is visually observed (Coad et al., 2014). These programs have limited capacity for environmental managers to adequately monitors and respond to algal blooms due to constraints such as (i) the expense of field monitoring, (ii) staffs availability and resources, (iii) field safety issues, and (iv) large time intervals between data collection, reporting and public notification. Therefore, to decrease the cost of aquatic environmental in-situ monitoring and increase the accuracy of blooms forecasting, an early-warning proactive approach of the algae blooms forecasting model is essential to prevent or mitigate the occurrence of algal blooms, and eventually facilitate the minimization of the adverse effect of algal blooms on the water bodies (Oh et al., 2007).

In terms of the model development, there are typically two kinds of forecasting models: deductive models and inductive models. The deductive models are developed based on the existing theories and knowledge which enable users to simulate the systems' behavior (Recknagel, 1997). Many deductive models have been proposed to predict the algal blooms. For example, Grover (1991) built a mechanistic deductive model by considering the theoretical elements needed for algae growth. Wei et al. (2014) used a coupled hydrodynamic-algal biomass model to forecast the short-term cyanobacteria blooms in Taihu Lake, China. Except for the wide application, however, deductive models require detailed descriptions of physical, chemical and biological processes, and usually contain a lot of parameters for calibration (Zhang et al., 2014). To a great extent, therefore, the prediction accuracy was restricted by the lacking of knowledge about the algae-growth mechanism and chlorophyll dynamics.

In terms of the inductive models, they produce holistic information extracted from the empirical data patterns by statistic, correlation and machine learning methods which enable users to predict rather than to explain the systems' behavior (Recknagel,

* Corresponding authors.
E-mail addresses: zl_liao@tongji.edu.cn (Z. Liao), jin.zhang@hotmail.com (J. Zhang).

http://dx.doi.org/10.1016/j.ecolmodel.2017.09.013
0304-3800/

1997; Zhao et al., 2016). Jeong et al. (2003) modeled microcystins aeruginosin bloom dynamics in Nakdong River by means of evolutionary computation and statistical approach. Cha et al. (2014) reported a Bayesian Poisson model to probabilistically predict the *Cyanobacteria* abundance in a Korean reservoir.

Among these inductive models, artificial neural network (ANN) is a kind of model which could deal with the complex information among the data by using machine learning theories. For normal forward neural networks, their structures include input-layer, hidden-layer, and output-layer. There are multiple neurons in each layer, and each of them is assigned with two parameters, namely weight value and the threshold value. By training ANN with enough samples obtained from the history of a natural system, these parameters can be adjusted to reproduce the behavior of this system (Ethem, 2010). ANN has been applied to the algae blooms and chlorophyll dynamics forecasting in the last two decades. For example, ANN was used for modeling and prediction of algal blooms (Recknagel et al., 1997) and also for modeling and prediction of zooplankton dynamics in Lake Kasumigaura (Recknagel et al., 1998). In Huelva, Western Andalucia, Spain, ANN was adapted for one-step weekly prediction of *Dinophysis Acuminata* blooms (Velo-Suárez and Gutiérrez-Estrada, 2007).

The ANN model was also applied on chlorophyll dynamics, as it is one of the factors to represent some species of algae and has been regarded as one of the early-warning proactive approaches to prevent or mitigate the occurrence of some algal blooms. To improve the understanding of chlorophyll dynamics, Coad et al. (2014) used ANN to predict daily *Chlorophyll-a* concentration. In a case study on the Yuqiao Reservoir in North China, an ANN was employed for the eutrophication forecasting and management (Zhang et al., 2015). Despite its successful application, it is stressed that the optimal ANN is generally problem dependent (Maier and Dandy, 2000; Recknagel, 2001). For this reason, it is necessary to develop and optimize the ANN for different problems to obtain the best model configurations that have a lower error with short training time and higher accuracy. Traditionally, an optimal ANN model was found by trial and error with adjusting of its structures and parameters (Maier and Dandy, 2001; Dedecker et al., 2004). However, it is difficult to find the optimal set of the possible structures and parameters. Therefore, a new method for developing and optimizing ANN models with the easier operation to predict chlorophyll dynamics is needed.

In terms of the chlorophyll dynamics, it has been widely recognized that the water quality, hydrology, and climate condition are the main influencing factors on the chlorophyll dynamics (Coad et al., 2014; Seitzinger, 1991). Compared with the base value of chlorophyll (refers to the value at the beginning of a period used as a reference or starting point for the estimation process), the value change of chlorophyll (refers to the difference between the size of the value to the end and the beginning of a period) is more sensitive to these influencing factors. In other words, the influencing factors will first determine the value change of chlorophyll, and then influence the base value of chlorophyll within a given time period. Thus, the primary hypothesis of this study is that the ANN based chlorophyll dynamics prediction model could be optimized by computing the correlations between the value change of chlorophyll and its influencing factors rather than the correlations between the base value and the influencing factors.

Consequently, the objectives of this study were to (i) explore a method to optimize ANN based chlorophyll dynamics prediction model, (ii) apply this optimized model and forecast daily *Chlorophyll-a* concentrations in a water body, and (iii) compare the accuracy of the optimized model with a traditional ANN based chlorophyll dynamics prediction model.

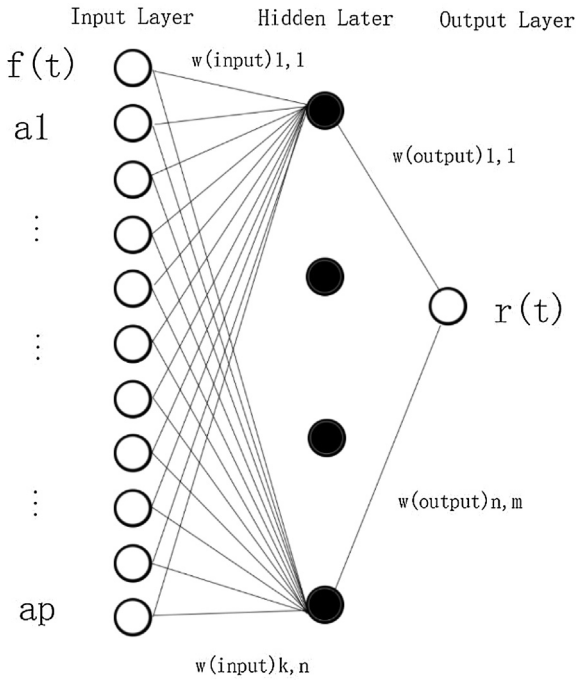

Fig. 1. Structure of an ANN model.

2. Materials and methods

2.1. Theory of ANN model

ANN applied in this study consists of an input layer with *p+1* nodes, a hidden layer with N nodes, and an output layer with one node as given in Fig. 1. *p* is the number of variables which depends on different the model designing. The number N is subjected by Eq. (1) as follows.

$$N \le \frac{N^{TR}}{\left(N^{I}+1\right)} \tag{1}$$

where, N is the nodes number in the hidden layer; N^{TR} is the number of training samples; and N^{I} is the number of inputs. There are mainly two kinds of transfer functions for each node, i.e.: the log sigmoid as shown in Eq. (2) and the tangent sigmoid as shown in Eq. (3). In this study, the tangent sigmoid was chosen as the transfer function for the hidden layer nodes and the output layer nodes.

$$\varphi_0 = \log sig\,(x) = \frac{1}{1+exp\,(-x)} \tag{2}$$

$$\varphi_h = \tan sig\,(x) = \frac{2}{1+exp\,(-2x)} - 1 \tag{3}$$

The connections among nodes in each layer are represented by the weights ($W\,(input)$ and $W\,(output)$) and thresholds (b_1 and b_2). The initial values of weights are determined by a random starting, and randomly set between −1 and 1. Thresholds are corresponding with inputs which are set randomly between −1 and 1 at the beginning. Then the weights and thresholds are adjusted by training with samples. If let functions $a_i\,(t)$, $i = 1, \ldots, p$ as the inputs of ANN, and let function $f\,(t)$ as the output of ANN, t is time variable. Then the model is represented by the following equations:

$$f(t)_N = \varphi_h\{\sum_{j=1}^{N} W(output)_j \cdot \varphi_h[\sum_{i=1}^{p} W(input)_{ij} \cdot I_{iN}\,(t) + b_{1j}] + b_2\} \tag{4}$$

$$I_{iN}(t) = 2\frac{a_i(t) - a_{imin}(t)}{a_{imax}(t) - a_{imin}(t)} - 1, t_0 \le t < t_k \quad (5)$$

$$f(t)_N = 2\frac{f(t) - f(t)_{min}}{f(t)_{max} - f(t)_{min}} - 1, t_0 \le t < t_k \quad (6)$$

where φ_0 is the activation function for the output layer given by Eq. (2); φ_h is the activation function for the hidden layer given by Eq. (3); $I_{iN}(t)$ is the normalized operation for input $a_i(t)$ given by Eq. (5);. and $a_i(t)$, i = 1, . . ., p are the inputs of ANN which influencing the $f(t)$. Furthermore, Eq. (6) is the normalized operation for the chlorophyll value $f(t)$. And the model can be given as Eq. (4).

Each data of the samples $a_i(t)$ was firstly normalized as $I_{iN}(t)$, because the values of inputs and output ranged up to different level to reduce the training effect of ANN. $a_{imax}(t)$ and $a_{imin}(t)$ are the maximum and minimum of$a_i(t)$ in the time interval $[t_0, t)$, which is the same for $f(t)_{max}$, $andf(t)_{min}$.

2.2. Optimization of the traditional ANN model

Traditional ANN models for chlorophyll prediction normally attempted to figure out the relationship between the value of chlorophyll and its influencing factors. For instances, Velo-Suárez and Gutiérrez-Estrada (2007) used weekly *Dinophysis acuminata* concentrations (cells/L) as input variables of ANN to predict the growth trends in blooms population dynamics. Coad et al. (2014) used 15 min intervals *Chlorophyll-a* concentrations (μg/L) and other parameters to configure an ANN to predict (one, three and seven days in advance) the mean, 10th and 90th percentile, and daily *Chlorophyll-a* concentrations. The main hypothesis about the optimization of the traditional ANN model is attempted to figure out the relationship between the change of value and its influencing factors.

Consequently, let $f(t)$ refers to the value of chlorophyll at time t. Usually, $f(t)$ is a function of the environmental factors which influence the chlorophyll dynamics. Then, the traditional ANN is used to find the correlations between those factors and the base value of chlorophyll, and compute $f(t)$ directly. However, we hypothesize that the traditional ANN models could be optimized by computing the value change of chlorophyll as flows.

Scatter time scale $[t_0, t_0 + T]$ into $\{t_k\}, k = 1, 2, \ldots n$, which subjects to:

$$t_0 = t_1 < t_2 < \ldots < t_{n-1} < t_n = t_0 + T$$

where t_0 is the beginning time; T is the time scale. Combining $\{t_k\}$ and $f(t)$ yields$\{f(t_k)\}, k = 1, 2, \ldots n$. If n is large enough, then $\{f(t_k)\}$ almost equals $f(t)$ within$[t_0, t_0 + T]$. $\{f(t_k)\}, k = 1, 2, \ldots n$, can be further generated by following iterated algorithm:

$$f(t_{k+1}) = f(t_k) + \Delta(t_k)\, k = 2, 3, \ldots, n \quad (7)$$

where $\Delta(t_k)$ is the value change of chlorophyll during the time scale from t_k tot_{k+1}. In addition, $\Delta(t_k)$ is a function of the initial chlorophyll population $f(t_k)$ and the other influencing factors. The i^{th} influencing factor is defined as a_i(t), i = 1, 2, . . ., p. Then, the function of $\Delta(t_k)$ is expressed as:

$$\Delta(t_k) = \Delta\left[a_1(t_k), \ldots, a_p(t_k), f(t_k)\right] \quad (8)$$

Since the relationship between $\Delta(t_k)$ and the influencing factors is too complex to calculate, and $\{\Delta(t_k)\}, k = 2, \ldots, n$ cannot be solved directly, ANN is employed. Then, the ANN for calculating $\Delta(t_k)$ is given as follows:

$$\Delta(t_k)_N = \varphi_h\{\sum_{j=1}^{N} W(output)_j \cdot \varphi_h[\sum_{i=1}^{p} W(input)_{ij} \cdot I_{iN}(t_k) + b_{1j} + f(t_k)_N \cdot W(input)_{(p+1)j} + b_{1(p+1)}] + b_2\} \quad (9)$$

$$\varphi_h = \tan sig(x) = \frac{2}{1 + exp(-2x)} - 1 \quad (10)$$

$$I_{iN}(t_k) = 2\frac{a_i(t_k) - a_{imin}(t)}{a_{imax}(t) - a_{imin}(t)} - 1, t_0 \le t < t_k \quad (11)$$

$$f(t_k)_N = 2\frac{f(t_k) - f(t)_{min}}{f(t)_{max} - f(t)_{min}} - 1, t_0 \le t < t_k \quad (12)$$

$$\Delta(t_k)_N = 2\frac{\Delta(t_k) - \Delta(t)_{min}}{\Delta(t)_{max} - \Delta(t)_{min}} - 1, t_0 \le t < t_k \quad (13)$$

$a_{imax}(t)$ and $a_{imin}(t)$ are the maximum and minimum of $a_i(t)$ in the time interval $[t_0, t)$, which is the same for $\Delta(t)_{max}$, $\Delta(t)_{min}$, $f(t)_{max}$, $andf(t)_{min}$.

2.3. Indices of the model performance

Four indices are used to evaluate the performances of networks' training and prediction. The selected performance indices are the mean square error (*MSE*, Eq. (14)), *R* (Eq. (15), Eqs. (16) and (17)), bias feature (*BF*, Eq. (18)) and accuracy feature (*AF*, Eq. (19)).

$$MSE = \frac{1}{n}\sum_{k=1}^{n} |\left(f(t_k) - f_p(t_k)\right)|^2 \quad (14)$$

$$R = \sqrt{1 - \frac{F}{F_0}} \quad (15)$$

$$F = \sum_{k=1}^{n}\left[f(t_k) - f_p(t_k)\right]^2 \quad (16)$$

$$F_0 = \sum_{k=1}^{n}[f(t_k) - f(t)_{meanobserved}]^2 \quad (17)$$

$$\log(BF) = \frac{\sum_{k=1}^{n} log\frac{f_p(t_k)}{f(t_k)}}{n} \quad (18)$$

$$\log(AF) = \frac{\sum_{k=1}^{n} |log\frac{f_p(t_k)}{f(t_k)}|}{n} \quad (19)$$

where n is the total number of the scattered time point; $f_p(t_k)$ and $f(t_k)$ are the predicted and observed chlorophyll value at time t_k; $f(t)_{meanobserved}$ is the average value of observed chlorophyll value; *MSE* is normally used as measures of model accuracy. If *MSE* is close to 0, it indicates a very close approximation to the actual values. The *R* is a relative measure of fit. If the value is close to 1, it indicates a good performance. In addition, *BF* is a measure of systematic overestimation or underestimation of the target values; and *AF* characterizes the average difference between the observed and the predicted values. *AF* and *BF* values in the range from 0.95 to 1.11 indicate a model with a good performance (García-Camacho et al., 2016).

3. Model configuration

3.1. Study area and observed data

Both of the traditional and optimized models were applied to an estuary reservoir in East China for predicting the chlorophyll dynamics. The estuary reservoir has a surface area of about 60 km^2 with an effective storage capacity of about 4.35 hundred million m^3. High level of chlorophyll occur in this reservoir mainly resulted from excess nutrients coinciding shallow and slow-flowing water.

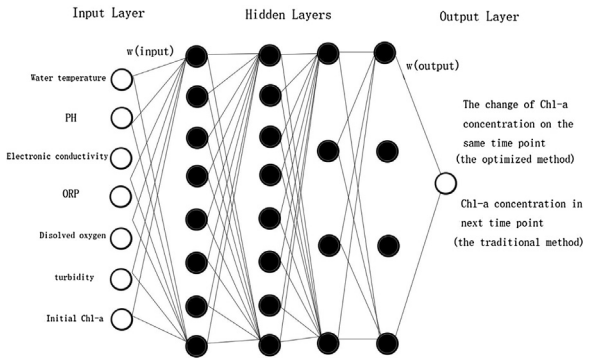

Fig. 2. Structures of the optimized and traditional ANN models.

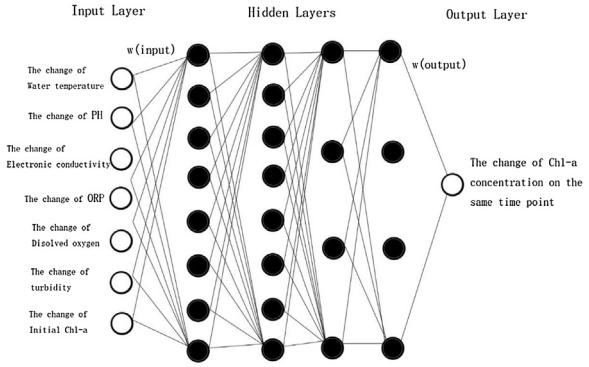

Fig. 3. Structure of the input-stabilized ANN model.

The water quality data was obtained from 5th to 9th June, and 24th to 26th July 2015. A total of 1152 samples were collected from the outlet of the reservoir with 10 min intervals. The parameters of temperature, pH, electronic conductivity, ORP (Oxidation-Reduction Potential), turbidity, DO (dissolved oxygen), and *Chlorophyll-a* were determined for each water sample.

3.2. Model configuration

3.2.1. Configuration of ANN model

The observed data were presented as the input variables of the optimized and tradition models. The concentration change of *Chlorophyll-a* on the same time point was presented as the output variable of the optimized model. The total amount of *Chlorophyll-a* concentration on the next time point was presented as the output variable of the traditional model. All the networks had 7 nodes in the input layer and 1 node in the output layer. To optimize the structures of all ANNs, it is necessary to initially decide their numbers of nodes in the hidden layers. According to Eq. (1), the node number in hidden layer has the limitation. In this study, $N^{TR} = 288$ as there were 882 training data; $N^I = 7$, as there were 7 influencing factors presented to the input layer. Consequently, the upper limit number of nodes in the hidden layers were 8 ($N^{TR}/\left(N^I + 1\right) = 288/8 \approx 36$. Therefore, $N \leq 36$). To further improve the generalization ability of ANN, four hidden layers with 24 nodes ($24 \leq 36$) were designed and given in Fig. 2

3.2.2. Input-stabilized model

To render input variables consistent with output variables, the optimized ANN model was further adapted by the input stabilization. In order words, an input-stabilized model was established which used the changes of observed data as input variables. The changes of the observed data between two time-points were presented as the input variables of the input stabilized optimized model. The concentration changes of *Chlorophyll-a* on the same time point was presented as the output variable of the input-stabilized model. As $N^{TR} = 288$; $N^I = 7$, the nodes in the hidden layers should less than 25 according to Eq. (1), which is the same as the optimized model. To further improve the generalization ability, four hidden layers with 24 nodes ($24 \leq 36$) were chosen for the input-stabilized model and was given in Fig. 3

3.2.3. Training algorithm, early stopping technique and data set

The back propagation algorithm was used for training the model (Rumelhart et al., 1985). To avoid overfitting and make sure the model's generalization ability can be improved by the back propagation algorithm, Early Stopping Technique (EST) was used. Based on EST, samples had been separated into three parts: train, validation, and test sets in the training process. The train set was used to train the model. When each epoch of training was finished, the algorithm would send a sample from the validation set into the model to compute *MSE*. If the *MSE* of the validation set no longer decreases for some steps (validation checks in the training process), the algorithm would stop the training process to avoid the overfitting. Finally, the trained model would be used to predict the test set to check its generalization ability. The train set *R* represents the ability to fit the train set; the validation *R*, and test *R* estimate the generalization ability. It would show a high possibility of overfitting and lack generalization ability if there is a significant difference between the *R* values of validation and test sets.

MSE versus the number of the epoch was given for training, validation and testing runs for all models. In the training phase, 288 samples from 5th June to 6th June 2015 were used as train, validation and test sets for training the first models. And 576 samples from 5th June to 6th June 2015 and 22nd to 23rd June 2015 were used as train, validation and test sets for training the second model. In the predicting phase, the data from 7th June to 9th June 2015 were used to assess the performance of the first network, and the data from 24th June to 26th June (864 samples) were used to assess the performance of the second network. The networks training, testing, validation and predicting were conducted with the MATLAB platform R2012b.

4. Results

4.1. Results of training phase

4.1.1. Training phase of first network

According to Fig. 4(a)–(c), the *MSE* values for the train, validation and test sets of all the network ranged from 1.01 to 0.4 after 100 epochs. The results indicated that all models could fit all the training, validation and test sets and achieve the satisfactory convergences. The gradient of all the networks went down and reach a low value at the end of epochs (the gradient was 3.087 at epochs 218 for the optimized model, 1.3398 at epochs 590 for the traditional model, 0.05576 at epochs 337 for the input-stabilized optimized model), which indicated that the back propagation algorithm worked well in the training runs. Val fail shows the validation counter in each training epochs, if the *MSE* of the validation set in the training samples did not decrease compared to the former epoch, the counter would plus 1; else it would be reset as 0 and send to the next epoch. If the counter keeps increasing to the validation check, the algorithm would stop the training process to avoid overfitting. The validation checks of all the models were 100 at the end of an epoch which means all the networks had little probability of overfitting. Also, the learning rate in each epoch of the training process for all the networks was given in Fig. 5. All the learning rate in each epochs were in the range of 0–0.4, which is acceptability during the training process.

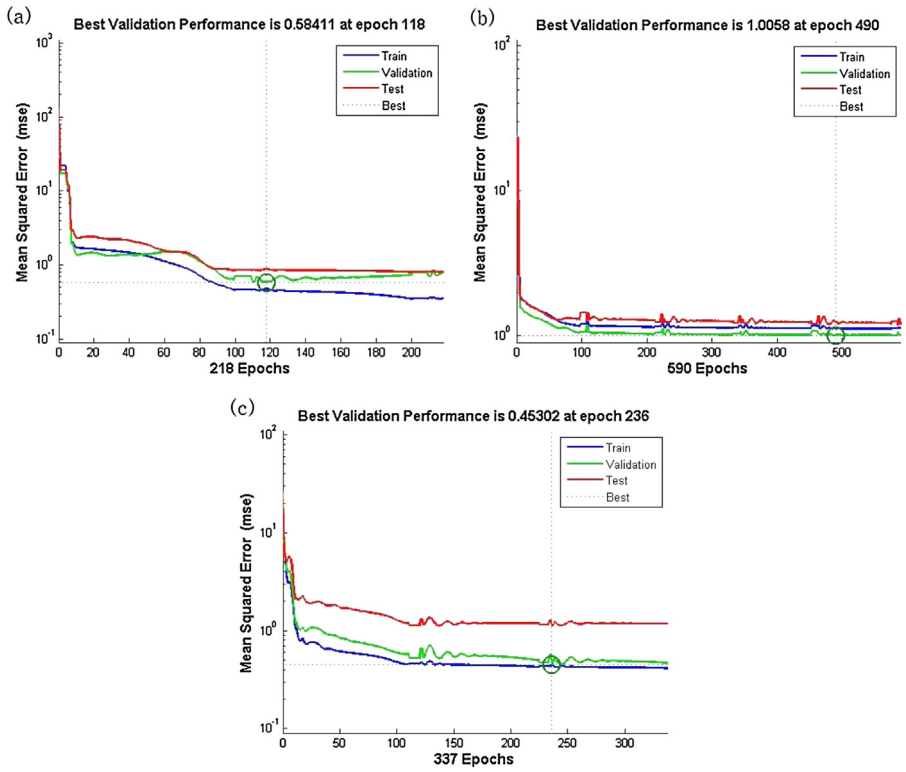

Fig. 4. *MSE* versus epochs for (a) the optimized, (b) the traditional, and (c) the input-stabilized optimized models.

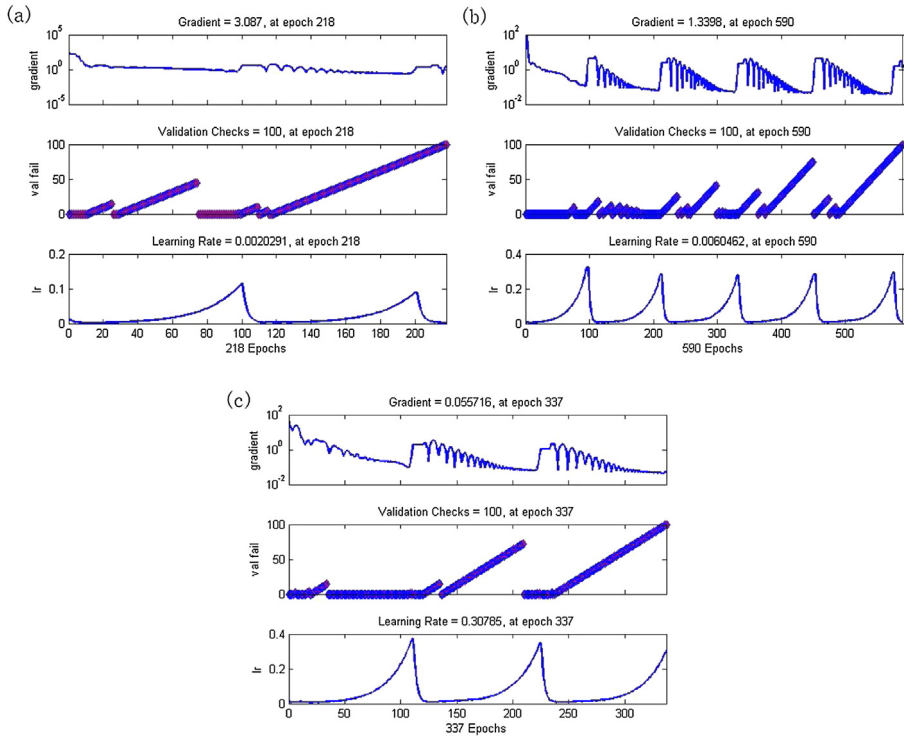

Fig. 5. Gradient, validation checks and learning rate information with epochs for (a) the optimized, (b) traditional, and (c) the input-stabilized optimized models.

Regression results of *R* in train, validation and test sets for all the models were given in Fig. 6. It shows that all the *R* values of the optimized models were higher than the corresponding *R* of the traditional model (0.86831 and 0.87254 compared with 0.65183; 0.81156 and 0.88513 compared with 0.66597; and 0.83348 and 0.76945 compared with 0.65589). It suggests that the optimized and the input-stabilized optimized network performed better than the traditional network in the train set fitting and more likely to have a better generalization ability based on the same data set.

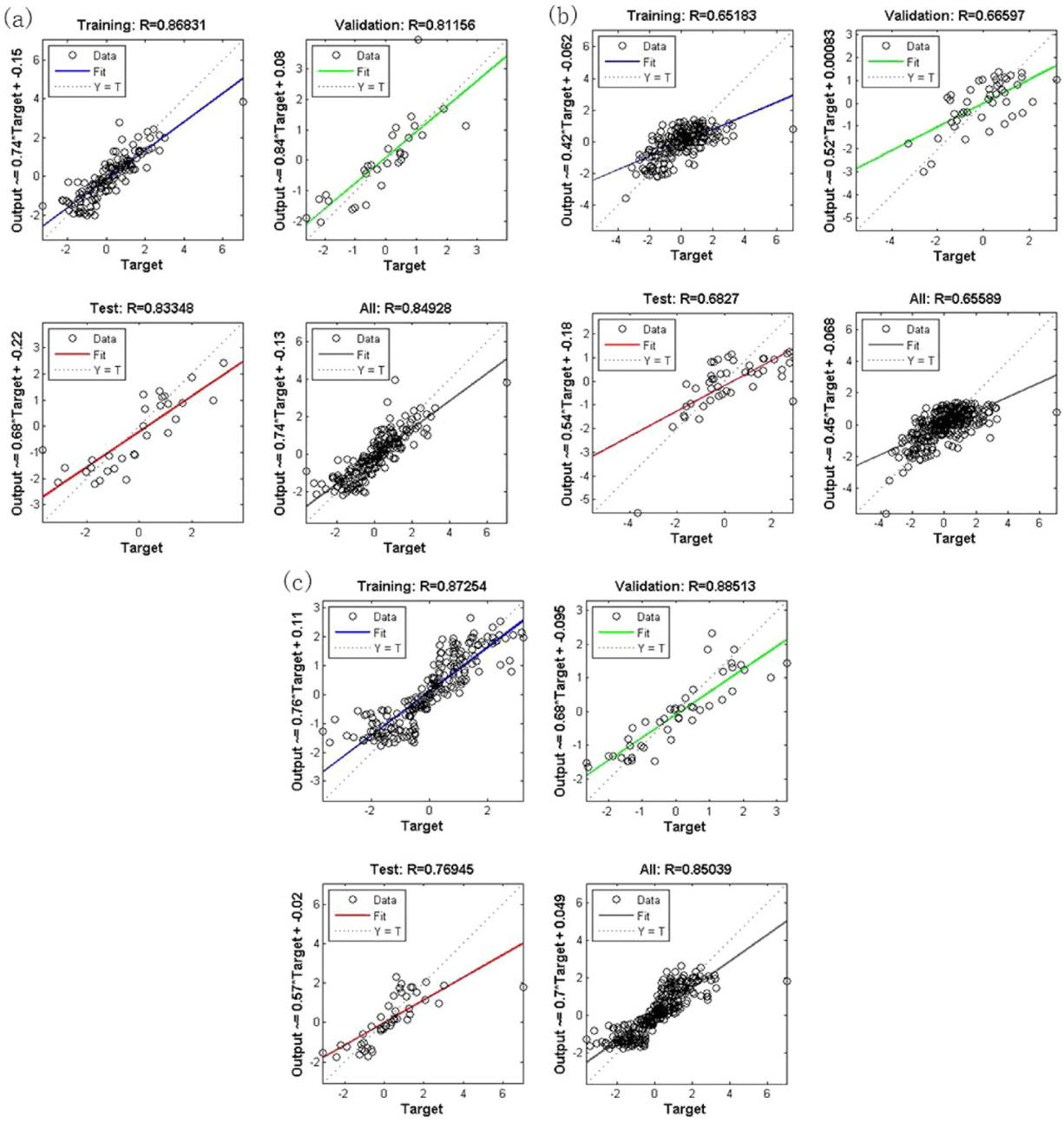

Fig. 6. Regression results of (a) the optimized, (b) traditional, and (c) the input-stabilized optimized models.

4.1.2. Training phase of second network

In the phase of training, the MSE values for the train, validation and test sets of all the network ranged from 1.01 to 0.2 after 300 epochs (according to Fig. 7(a), (b), and (c)), which that all models could fit all the training, validation and test sets and achieve the satisfactory convergences. The back propagation algorithm worked well in the training runs as the gradient of all the networks went down and reach a low value at the end of epochs (the gradient was 0.106 at epochs 397 for the optimized model, 0.4046 at epochs 156 for the traditional model, 0.0967 at epochs 190 for the input-stabilized optimized model). The validation checks of all the models were 100 at the end of an epoch which means all the networks had little probability of overfitting. And all the learning rate in each epochs were in the range of 0 to 0.5 (given in Fig. 8) which is acceptability during the training process.

Regression results of R in train, validation and test sets for all the models were given in Fig. 9 (0.85565 and 0.89518 compared with 0.55561; 0.87422 and 0.85722 compared with 0.66063; and 0.85267 and 0.8725 compared with 0.56887). It suggests that both of the optimized network performed better than the traditional network in the train set fitting and more likely to have a better generalization ability based on the same data set, which was same to the first training phase.

4.2. Results of prediction phase

4.2.1. Prediction on 7th–9th June 2015

The predicting values on 7th to 9th June 2015 of all the models compared with the observed values were given in Fig. 10. For the optimized models, results were calculated based on Eq. (7), which means that the values given were calculated by adding

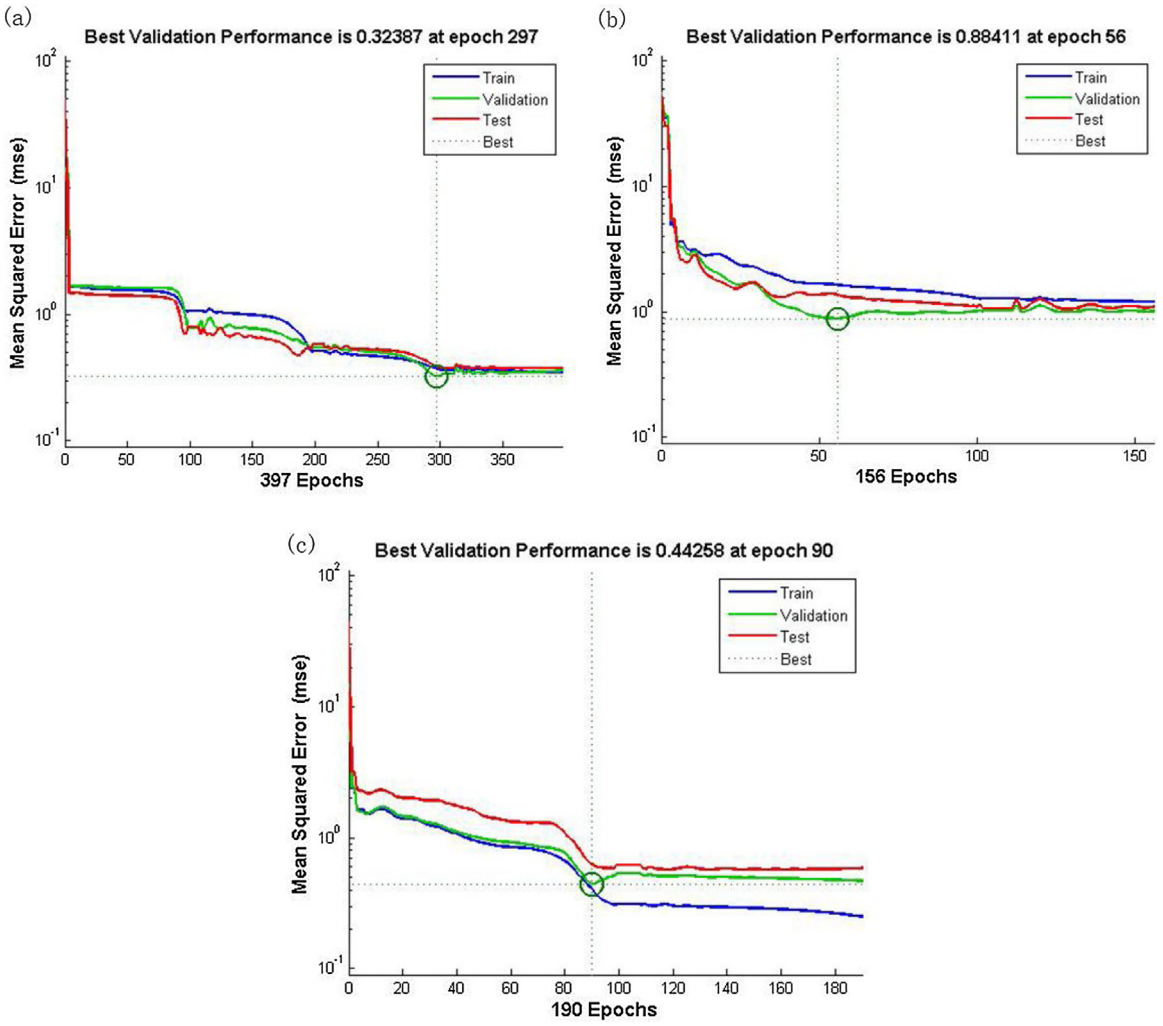

Fig. 7. *MSE* versus epochs for (a) the optimized, (b) the traditional, and (c) the input-stabilized optimized models.

Table 1
The *MSE*, *BF*, *R* and significance *P* (*T*-test) of all the models in predicting.

	Optimized model	Traditional model	Input-stabilized model
MSE	1.0161	1.1671	0.5143
BF	0.9847	0.9848	1.0019
AF	1.0751	1.0741	1.0310
R	0.8302	0.7507	0.9546
Significance *P*	0.7266	2.13e-5	0.6026

Table 2
The *MSE*, *BF*, *R* and significance *P* (*T*-test) of all the models in predicting.

	Optimized model	Traditional model	Input-stabilized model
MSE	1.3349	2.0422	0.76714
BF	0.9869	0.9733	1.0042
AF	1.0594	1.0962	1.0335
R	0.7649	0.6742	0.8986
Significance *P*	0.0053	5.247e-10	0.6841

the initial *Chl-a* and the change *Chl-a* concentrations (ANN computed). It shows that the optimized models had better prediction results than the traditional model in reproducing the tendency of the observed values and in some certain specific points. As shown in Table 1, the values of predicting *MSE* of the optimized and traditional models were close to each other, which means that the general results of their predictions were similar. However, *MSE* of the input-stabilized model were lower than them of the optimized and traditional the models, which indicates a good general performance of the input-stabilized model. *R* of the optimized models were closer to 1 than the traditional model, indicating the optimized models have higher accuracy than the traditional model. Moreover, even the training process of both optimized models was not significantly different from each other, in the predicting process, *R* of the input-stabilized model was better than the optimized model. It suggests the input stabilized process can improve the generalization ability of the model. Significance *P* (T test) of the optimized and input-stabilized models were larger than 0.05 which means that there was no significant difference between the prediction of each model and the true value. Even *R* of the traditional model was acceptable. The *P* values of the traditional was lower than 0.05, which indicates a significant difference between the prediction of traditional model and true value. In addition, the *BF* values of the models were 0.9847 (optimized model), 0.9848 (traditional model), and 1.0019 (input-stabilized mode). The *AF* values of the models were 1.0751 (optimized model), 1.0741 (traditional model), and 1.0310 (input-stabilized mode). All the values were within the ranges of 0.95–1.11, which were considered to be acceptable (García-Camacho et al., 2016).

4.2.2. Prediction on 24th–26th June 2015

The predicting values on 24th–26th June 2015 of all the models compared with the observed values were given in Fig. 11. As shown in Table 2, the values of predicting *MSE* of the optimized

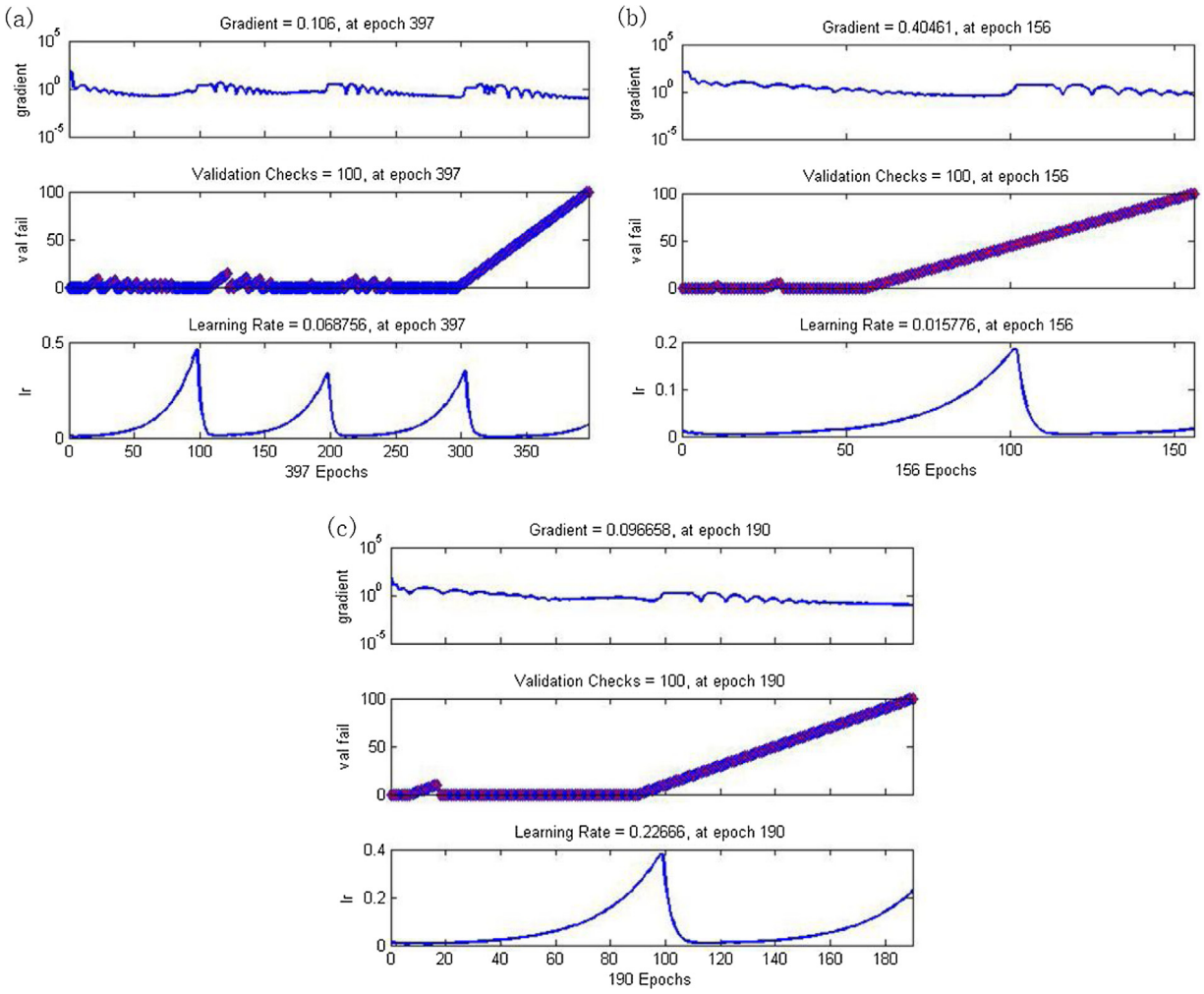

Fig. 8. Gradient, validation checks and learning rate information with epochs for (a) the optimized, (b) traditional, and (c) the input-stabilized optimized models.

and input-stabilized models were better than the traditional one, which means that the general results of the optimized and input-stabilized predictions were better. As *R* of the optimized models were closer to 1 than that of the traditional model, the optimized models had better results in prediction. Similar to the first test, the *R* of the input-stabilized model was better than the optimized model, indicating a better generalization ability. Significance *P* (T student test) of the optimized and input-stabilized models were larger than 0.05 which means that there was no significant difference between the prediction of each model and true value. However, similar to the first test, the *P* of the traditional model was lower than 0.05, which indicates a significant difference between the predictions of traditional model and true value. The *BF* values of all the models were 0.9869 (optimized), 0.9733 (traditional), and 1.0042 (input-stabilized). The *AF* values of all the models were 1.0594 (optimized), 1.0962 (traditional), and 1.0335 (input-stabilized). All the values were within the ranges considered to be acceptable. In general, even all the results of the second test were not as good as the first test, the optimized and input-stabilized models were better than the traditional one, which was consistent with the first test results.

Compared with the first test, the prediction from 24th to 26th June were not as good as the former one. The possible reason was that the time delay of data (the data from 5th June to 6th June 2015 in the training samples) period for training and prediction influence the model performance. So that an extra test was applied here to figure this. All the three models were trained only by the data from 22nd to 23rd June 2015 in the extra test, and used to predict Chl-a values on 24th to 26th June 2015. All the results compared with the observed values were given in Fig. 12. And the *MSE*, *BF*, *AF*, *R* and Significance *P* were given in Table 3. It shows that the performance of extra test is little bit better than the second prediction, which indicates that the model performance was influenced by the time delay data period for training and prediction.

Table 3
The MSE, BF, R and significance P (T-test) of all the models in the extra test.

	Optimized model	Traditional model	Input-stabilized model
MSE	1.303	1.8992	0.6586
BF	0.9894	0.9354	1.0005
AF	1.0567	1.0894	1.0277
R	0.774	0.6838	0.9076
Significance *P*	0.1791	2.12E-19	0.9495

5. Discussion

Both of the optimized and traditional networks had been applied to the case study. According to the results, the optimized model had a better prediction ability than the traditional model, which confirmed our hypothesis that the change of chlorophyll value is more sensitive to the influencing factors than the base value of chlorophyll. Furthermore, this phenomenon could be explained by the theory of non-stationary time series. A non-stationary time series could be defined as the statistical characteristics change over

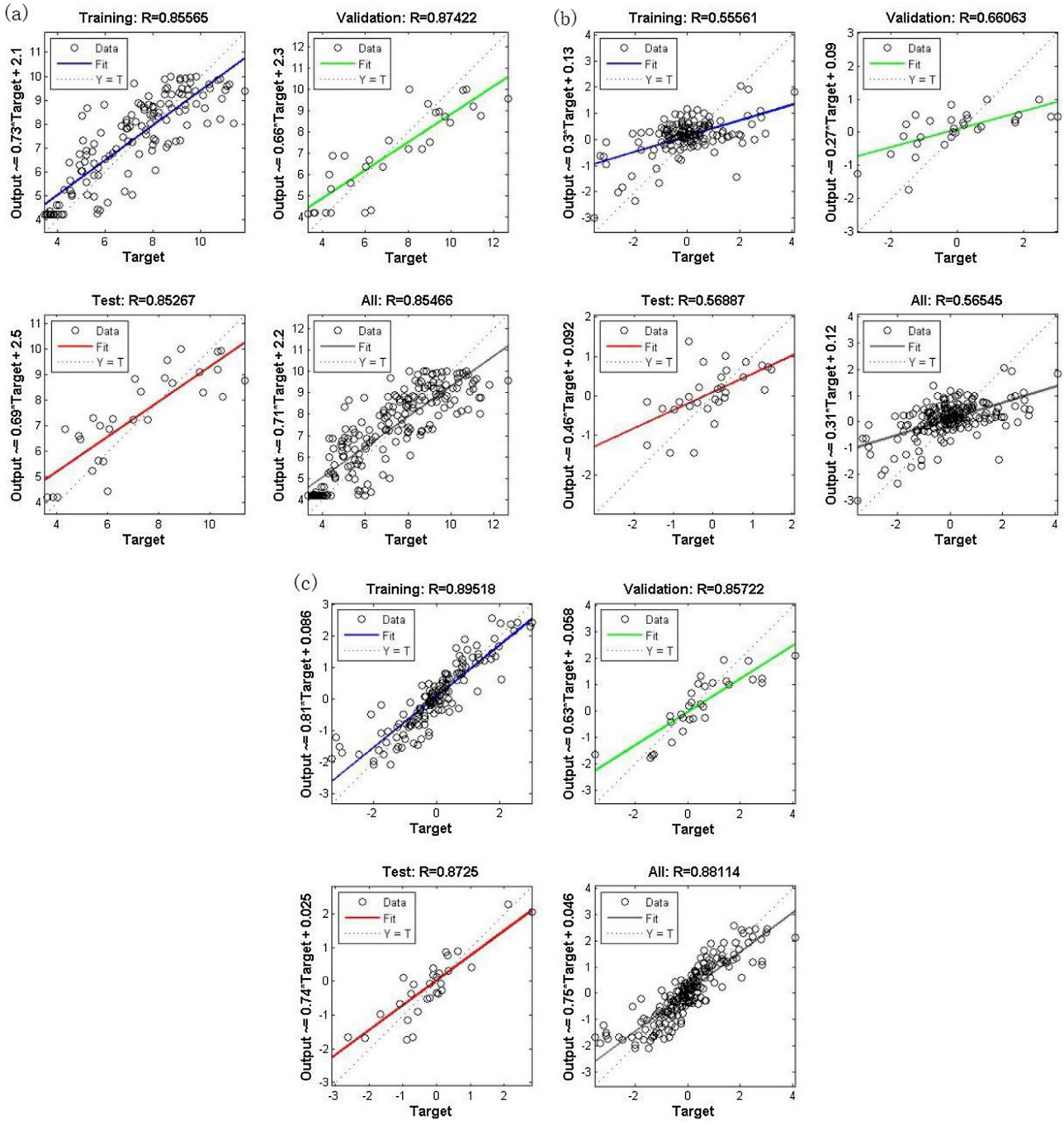

Fig. 9. Regression results of (a) the optimized, (b) traditional, and (c) the input-stabilized optimized models.

time due to either internal or external dynamics (Maheswaran and Khosa, 2015). Therefore, the chlorophyll value series at each time is a non-stationary time series process, which changes over time due to the complex environmental influencing factors. These complex environmental influencing factors are major obstacles to the chlorophyll dynamics forecasting.

As a pre-processing or fundamental approach of non-stationary time series, differencing can stabilize a time series by removing changes in the level of a time series and eliminating trend and seasonality. It computes the differences between the consecutive observations and uses the differences to replace the original data for modeling and analysis. More explicit, let $\{f(t_k)\}, k = 1, 2, \ldots n$ be a non-stationary time series. $\{f(t_k)\}$ can be transformed to a stationary time series by differencing, i.e., $\{f(t_{k+1}) - f(t_k)\}, k = 1, 2, \ldots n-1$, which is more stationary than $\{f(t_k)\}$ with lower random influences in the data for modeling.

Also, the input-stabilized process can further improve the prediction results by improving the generalization ability of ANN. It could be explained by the fact that the stabilization on the input can make the surface of the parameter set more smooth, which is more suitable for gradient-based training algorithm to find an optimal solution and helps the model improve its generalization ability.

Comparing the second prediction and the extra prediction, it shows that the time delay of data period for training and prediction

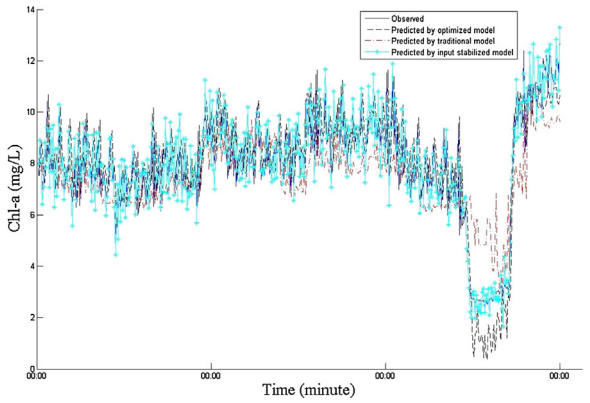

Fig. 10. Prediction results of all the models.

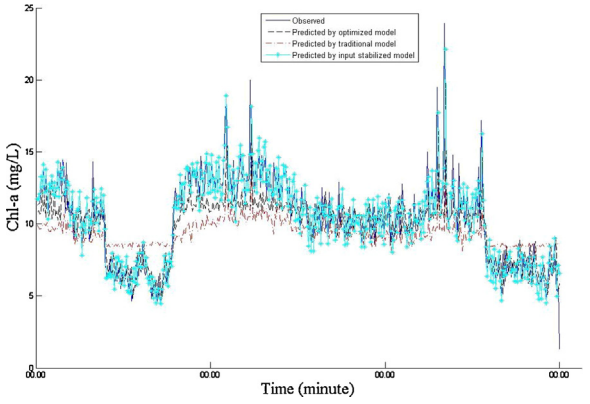

Fig. 11. Prediction results of all the models.

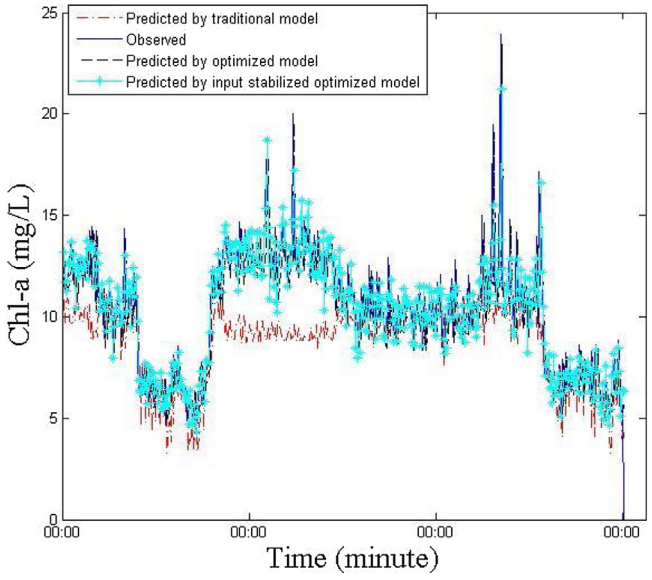

Fig. 12. Prediction results of all the models in the extra test.

influence the model performance. To further reduce the influence, more research should be considered.

Accordingly, the traditional ANN model was optimized by importing a differencing process, i.e., replacing the base value of chlorophyll with the change of chlorophyll value. Thus, the optimization approach proposed in this study was further confirmed by the concept of non-stationary time series from a theoretical aspect, which could be a universal method to improve the traditional ANN forecasting models in the other research.

6. Conclusions

In this study, a traditional artificial neural network (ANN) based chlorophyll dynamics prediction model had been optimized by presenting the change of chlorophyll value rather than the base value of chlorophyll as the output variable of the network. Both of the optimized and traditional networks had been applied to a case study. The results of model performance indices showed that the optimized network predicted better than the traditional network. Furthermore, this phenomenon had been explained from a theoretical aspect by means of the concept of non-stationary time series.

Consequently, the optimized network could be used as an early-warning proactive approach to predict the chlorophyll dynamics, and eventually facilitate the minimization of the adverse effect of chlorophyll dynamics on the water bodies.

Acknowledgements

This study was financially supported by National Natural Science Foundation of China (Grant No. 51578396), and German BMBF (Bundesministerium für Bildung und Forschung, Federal Ministry of Education and Research) CLIENT project “Managing Water Resources for Urban Catchments” in the framework of the Sino-German “Key Technologies and Management Modes for the Water Environmental Rehabilitation of a Lake City from the Catchment Viewpoint” (Grant No. 2016YFE0123300). The authors would like to gratefully acknowledge Mr. Haibing Shao and Prof. Olaf Kolditz for their valuable suggestions for this study.

References

Cha, Y.K., Park, S.S., Kim, K., Byeon, M., Stow, C.A., 2014. Probabilistic prediction of cyanobacteria abundance in a Korean reservoir using a Bayesian Poisson model. Water Resour. Res. 50 (3), 2518–2532.

Coad, P., Cathers, B., Ball, J.E., Kadluczka, R., 2014. Proactive management of estuarine algal blooms using an automated monitoring buoy coupled with an artificial neural network. Environ. Modell. Softw. 61 (C), 393–409.

Dedecker, A.P., Goethals, P.L.M., Gabriels, W., Pauw, N.D., 2004. 2004. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). Ecol. Modell. 174 (1), 161–173.

Ethem, A., 2010. Introduction to Machine Learning, Second ed. Adaptive Computation & Machine Learning, America.

García-Camacho, F., López-Rosales, L., Sánchez-Mirón, A., Belarbi, E.H., Chisti, Y., Molina-Grima, E., 2016. Artificial neural network modeling for predicting the growth of the microalga karlodinium veneficum. Algal Res. 14, 58–64.

Grover, J.P., 1991. Resource competition in a variable environment: phytoplankton growing according to the variable-internal-stores model. Am. Nat. 1991, 811–835.

Gu, X., Liao, Z., Zhang, G., Xie, J., Zhang, J., 2017. Modelling the effects of water diversion and combined sewer overflow on urban inland river quality. Environ. Sci. Pollut. Res., 1–12.

Jeong, K.S., Kim, D.K., Whigham, P., Joo, G.J., 2003. Modelling Microcystis aeruginosa bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. Ecol. Modell. 161 (s1–s2), 67–78.

Maheswaran, R., Khosa, R., 2015. Wavelet Volterra Coupled Models for forecasting of nonlinear and non-stationary time series. Neurocomputing 149, 1074–1084.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Modell. Softw. 15 (1), 101–124.

Maier, H.R., Dandy, G.C., 2001. Neural network based modelling of environmental variables: a systematic approach. Math. Comput. Modell. 33 (6–7), 669–682.

Oh, H.M., Ahn, C., Lee, Y., Chon, J.W., Choi, T.S., 2007. Community patterning and identification of predominant factors in algal bloom in Daechung Reservoir (Korea) using artificial neural networks. Ecol. Modell. 203 (s1–s2), 109–118.

Recknagel, F., French, M., Harkonen, P., Yabunaka, K.I., 1997. Artificial neural network approach for modelling and prediction of algal blooms. Ecol. Modell. 96 (s1–s3), 11–28.

Recknagel, F., Fukushima, T., Hanazato, T., Takamura, N., Wilson, H., 1998. Modelling and prediction of phyto- and zooplankton dynamics in Lake

Kasumigaura by artificial neural networks. Lakes Reservoirs Res. Manag. 3 (2), 123–133.
Recknagel, F., 1997. ANNA –Artificial Neural Network model for predicting species abundance and succession of blue-green algae. Hydrobiologia 349 (1), 47–57.
Recknagel, F., 2001. Applications of machine learning to ecological modelling. Ecol. Modell. 146 (s1–s3), 303–310.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1985. Chapter 8, learning internal representation by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1. MIT Press, Cambridge, MA, pp. 318–362.
Seitzinger, S.P., 1991. The effect of pH on the release of phosphorus from Potomac estuary sediments: implications for blue-green algal blooms. J. Estuarine Coastal Shelf Sci. 1991 33 (4), 409–418.
Velo-Suárez, L., Gutiérrez-Estrada, J.C., 2007. Artificial neural network approaches to one-step weekly prediction of Dinophysis acuminata blooms in Huelva (Western Andalucía, Spain). Harmful Algae 6 (3), 361–371.
Wei, L., Qin, B., Zhu, G., 2014. Forecasting short-term cyanobacterial blooms in Lake Taihu, China, using a coupled hydrodynamic-algal biomass model. Ecohydrology 7 (2), 794–802 (9).
Zhang, X., Recknagel, F., Chen, Q., Cao, H., Li, R., 2014. Spatially-explicit modelling and forecasting of cyanobacteria growth in Lake Taihu by evolutionary computation. Ecol. Modell. 306, 216–225.
Zhang, Y., Huang, J.J., Chen, L., Lan, Q., 2015. Eutrophication forecasting and management by artificial neural network: a case study at Yuqiao Reservoir in North China. J. Hydroinf. 17, 4.
Zhao, X., Du, K., Zhou, M., Ren, G., Li, C., 2016. Improvement of SVM regression forecast water supply model based on phase space reconstruction. J. Civil Architect. Environ. Eng. 38, 147–150.