Prediction and elucidation of algal dynamic variation in Gonghu Bay by using artificial neural networks and canonical correlation analysis

Heyi Wang
College of Hydrology and Water Resources
Hohai University
Nanjing, China
wangheyi@ciotc.org

Xuchang Yang
Bureau of Hydrology and Water Resources Monitoring Taihu
Basin Management Bureau
Wuxi, China
jennifer_why@qq.com

***Abstract*—This paper describes the training, validation and application of recurrent neural network (RNN) models to computing the algal dynamic variation at three sites in Gonghu Bay of Lake Taihu in summer. The input variables of Elman's RNN were selected by means of the canonical correspondence analysis (CCA) and Chl_a concentration as output variable. Sequentially, the conceptual models for Elman's RNN were established and the Elman models were trained and validated on daily data set. The values of Chl_a concentration computed by the models were closely related to their respective values measured at the three sites. The correlation coefficient (R^2) between the predicted Chl_a concentrations by the model and the observed value were 0.86-0.92. The results show that the CCA can efficiently ascertain appropriate input variables for Elman's RNN and the Elman's RNN can precisely forecast the Chl_a concentration at three different sites in Gonghu Bay of Lake Taihu in summer.**

***Keywords*—Elman's recurrent neural network; canonical correspondence analysis (CCA); Algal dynamic variation**

I. INTRODUCTION

Algal bloom is an environmental hazard which reduces the quality of water in rivers, lakes and reservoirs. Recurrent proliferation of algae usually causes a series of problems such as alteration of community structure, deterioration of water quality, and loss of cost-efficiency in water purification process [1-4]. Therefore there is a strong necessity for the establishment of appropriate ecological modelling systems, which can analyze behavior of proliferating phytoplankton with high accuracy.

In view of the complexity of aquatic food webs and their interactions with environmental variables, recurrent neural network (RNN) models capable of modelling a complex nonlinear system are required to elucidate and predict underlying processes of algal blooms. Many researchers have used neural networks to simulate the timing and magnitude of algal blooms and to forecast the cyanobacteria abundance [5-7]. In order to process sufficient information from the target ecosystem or entity, the size of RNN models tended to become larger by applying diverse state variables. In an RNN, one of main task is to determine the model input variables that affect the output variable significantly. Many researches provided algorithms or paradigms for selecting suitable input variables [8,9].

In recent years multivariate statistical analysis, especially canonical correspondence analysis (CCA) has been widely employed to examine relationships in large-scale ecological data sets. Hansel-welch et al.[10]showed the annual variation in abundance of filamentous algae by using CCA. Ke et al. [11] used it to explore the phytoplankton succession during the spring-summer periods in 2004 and 2005 in Meiliang Bay of Lake Taihu. CCA was also performed to elucidate the relationship between Microcystis operational taxonomic unit composition and the environmental factors in Lake Taihu [12].

The present study utilized RNN and CCA for unraveling complex ecological relationships in the database of Gonghu Bay in summer, and forecasting of algae concentrations by means of water quality and meteorological data. The study aimed at: (1) elucidating the relationships between algal dynamic variation and environmental factors in Gonghu Bay by means of CCA (2) forecasting the algae concentrations and to different environmental variables by means of recurrent neural network.

II. STUDY SITE AND DATA

Gonghu Bay is located in the northeast of Lake Taihu which is the third largest freshwater lake in China, with an area of 146 km^2 and average depth of 2.0m. Until now, there are three main waterworks scattered along shore of Gonghu Bay and supplied approximately 0.7 billion m^3 drinking water annually from the lake to the surrounding cities, such as Suzhou and Wuxi. The blooms usually take place in June-October dominated by cyanobacteria in Gonghu Bay.

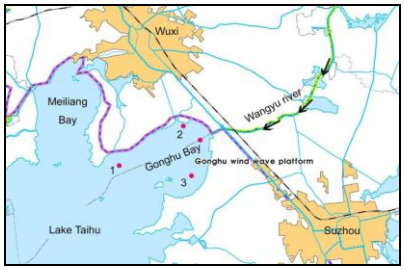

Fig. 1. Map showing the geographical setting of the present survey area with three sites

The data used in this study were derived from the water resource of Gonghu water quality survey and Gonghu wind-wave-platform conducted by Bureau of Hydrology and Water Resources Monitoring, Taihu Basin Management Bureau (TBA). The selected three sites (Fig.1) are designated as 1#, 2# and 3# closing three waterworks of Gonghu Bay. From June to October in 2010, sampling was undertaken once each day at three sites in the experiment area measuring environmental factors such as water temperature(WT, ℃), pH, dissolved oxygen(DO, mgL^{-1}), chemical oxygen demand (COD_{MN}, mgL^{-1}), total nitrogen(TN, mgL^{-1}), total phosphorus(TP, mgL^{-1}), Chlorophyll-a concentration(Chl_a, mgm^{-3}), the quantity of dilution water(WQ, m^3). Water samples were collected from a depth of 50cm below the surface, sampling network and analytical procedures are executed by standard.

TABLE I. THE QUANTITY OF DILUTION WATER DURING JULY TO OCTOBER IN 2010

Time	Total quantity (m^3)	Average input rate (m^3)
2010.6.30-2010.7.18	13169	693.1
2010.8.18-2010.9.11	21733	905.5
2010.10.11-2010.10.25	9754	650.3

The wind data collected at the several meteorological stations located around Gonghu Lake are similar [13]. Therefore, we used the wind data collected every minute at Gonghu wind wave platform (about 10m above the water level of Taihu Lake), which nearing site 2# (Fig.1). All winds were sampled at 1HZ and the wind measurements were conducted according to Marcel Bottema [15]. In this research, we choice 8 hours mean significant wave height before sampling as wave variable. And the wind-induced wave height H can be estimated using the following empirical formula [14]:

$$\frac{g\overline{H}}{U^2} = 0.22th\left(0.45\left(\frac{gd}{U^2}\right)^{0.72}\right)th\left(\frac{0.0016\left(gF/U^2\right)^{0.46}}{0.22th\left(0.45\left(gd/U^2\right)^{0.72}\right)}\right) \quad (1)$$

Where F and d represent the fetch and the depth of the site, U is the wind speed of 10m above the water level of Gonghu wind-wave platform, and g is gravity. Some statistics that describe the measured data, as well as the longitude and latitude for each sampling site are indicated TABLE Ⅰ and TABLE Ⅷ.

III. METHODS

A. The network structure

Elman's recurrent networks are a special type of the dynamic neural nets. The feedback connection in Elman's neural nets is from the outputs of neurons in the hidden layer to the context layer units that are called as context nodes. This part of the input layer, namely, the context layer, plays a role in storing internal states in Elman's net as mentioned above [16]. The result of processing in a previous time step can be used at the current time step. This property of the Elman type RNN provides very important advantage, especially, in real time applications to follow the dynamical change of water resources variables in practice.

In this study, three-layer Elman's neural networks were constructed for prediction of algal dynamic variation in three sites of Gonghu Bay, as shown in Fig.2. The model was composed of one input layer optimized input variables selected by the method of CCA, one hidden layer and one output layer with one output variable in three sites. In order to determine the optimum number of nodes in the hidden layer and transfer functions, different Elman models were constructed and tested.

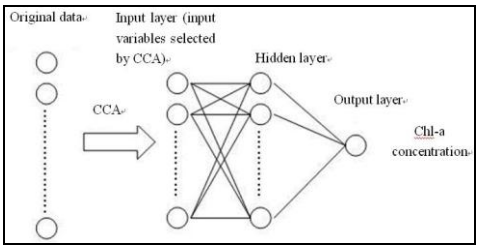

Fig. 2. The architecture of the Elman model for algae concentration in Gonghu Bay

B. Selection of input variables based on CCA

The choice of Elman's input variables is generally based on a priori knowledge of causal variables, inspections of time series plots, and statistical analysis of potential inputs and outputs. In this study, we applied CCA to determine the factors that influence the extent of the algal dynamic variation experienced by each environmental variable. We correlated the 8 environmental variables to algal dynamic variation. The goal of the method is to be able to select the algal dynamic variation best correlated with the environmental variables for the input variables of the Elman model subsequently. CCA creates pairs of linear combinations between each group of variables called canonical variables, so that the correlation between the variables of the same pare is maximized and so that correlation between the variables of two different pairs is nil. The analyses were performed with SAS 9 software. Details concerning CCA are available in reference books [17].

As a rule of thumb, an absolute value of 0.3 or greater in canonical loading was used to select the variables that are thought to have a meaningful interpretation of the related canonical variable [19, 20]. We chose a cutoff value of 0.45 to select important loadings in this study.

C. Model validation and neural network based sensitivity analysis approach

To determine the performance of selected network model, two different criteria were used: the Mean Relative Percentage Error (MRPE) and the coefficient of determination (R^2) [18]. The MRPE represents the error associated with the model and can be computed as:

$$MRPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - x_{pi}}{x_i}\right| \times 100 \quad (7)$$

The coefficient of determination (R^2) represents the percentage of variability that can be explained by the model and is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - x_{pi})^2}{\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i-1}^{n}x_{pi})^2} \quad (8)$$

Where x_{pi} and x_i represent the model computed and measured values of the variable, and N represents the number of observations. The MRPE, a measure of the goodness-of-fit, best describes an average measure of the error in predicting the dependent variable. Depending on sensitivity of algal dynamic variation and the mismatch between the forecasted algal dynamic variation and that measured; an expert can decide whether the predictability of the Elman model is accurate enough to make important decisions regarding data usage.

IV. Result and Discussion

A. Identification by canonical correlation analysis of the factors influencing Microcystis blooms CCA results

1) Site 1#

Only the first canonical correlations was statistically significant (F=1.83, P<0.001), indicating that the two sets of variables were correlated (TABLE Ⅱ). Axis 1 of the CCA, with the value of coefficient exceeds 0.78, explained 66.71% of the cumulative percentage variance of algal concentration while all the environmental variables considered for the analysis accounted for 89.21% of the total variance of the algal concentration.

TABLE II. Canonical correlation coefficients in site 1#

Axis	Axis1	Axis2	Axis3
Eigenvalues	1.8344	0.2488	0.0660
Species-environment correlations	0.7824	0.6348	0.3472
Cumulative (%)	66.71	76.53	89.21
Monte Carlo test: Eingenvalues-p	<.001	0.7081	0.1055

TABLE III. Canonical structures of the first pair of canonical variates in site 1# (CC=0.7824(Approx.F=1.83,P<0.001))

Environmental factors		Algal concentration	
variable	*loading*	*variable*	*loading*
WT(℃)	0.4686	chl_a_1	0.8085
pH	0.5759	chl_a_2	0.1210
DO (mgL^{-1})	0.2887	chl_a_3	0.2588
CODmn (mgL^{-1})	0.2717		
TN (mgL^{-1})	-0.1273		
TP (mgL^{-1})	0.3852		
8Hm(cm)	-0.6357		
WQ	-0.5486		

The canonical structures of the first pairs of canonical variates were shown in TABLE Ⅲ. It is shown that WT, pH, 8Hm and WQ were strongly correlated with the first CCA axis (0.4686, 0.5759, -0.6357 and -0.5486 respectively). This result helps to narrow down the relationship between the environmental factors and algal concentration. That is, hydrologic variables and wave climate might affect algal concentration in the next day. According the result of CCA and domain knowledge, WT, pH, 8hm and WQ were selected to the input variables of Elman model of Site 1#.

2) Site 2#

TABLE Ⅳ reveals that the first canonical correlation was statistically significant (F=5.63, P<0.001). The value of the first coefficient exceeds 0.92 and explained variance is about 84.7%. In fact, this value reflects a very strong link between the environmental factors and algal concentration. The canonical structures of the first pairs of canonical variates recorded in TABLE Ⅴ shows that the environmental factors mainly represented the hydrologic variables and wind-related wave height (WT, pH, 8Hm and WQ) strongly correlated to algal concentration mainly represented Chl_a_1. Obviously, hydrologic variables, wave climate and water division might affect algal concentration in the next day. According the result of CCA and domain knowledge, WT, pH, 8Hm and WQ were selected to the input variables of Elman model of Site 2#.

TABLE IV. Canonical correlation coefficients in site 2#

Axis	Axis1	Axis2	Axis3
Eigenvalues	5.6337	0.6750	0.2973
Species-environment correlations	0.9216	0.6348	0.4777
Cumulative (%)	84.70	91.14	100
Monte Carlo test: Eingenvalues-p	<.001	0.217	0.3222

TABLE V. Canonical structures of the first pair of canonical variates in site 2# (CC=0.9216(Approx.F=5.63,P<0.001)

Environmental factors		Algal concentration	
variable	*loading*	*variable*	*loading*
WT(℃)	0.4862	chl_a_1	0.9515
pH	0.4514	chl_a_2	0.3051
DO (mgL^{-1})	0.2420	chl_a_3	0.2300
CODmn (mgL^{-1})	0.3175		
TN (mgL^{-1})	-0.1582		
TP (mgL^{-1})	-0.4280		
8Hm(cm)	-0.6307		
WQ	-0.5528		

3) Site 3#

Based on the CCA results (TABLE Ⅵ), the first canonical correlations were statistically significant (F=1.66, P<0.001). The canonical correlation coefficient was 0.79. Altogether, these 8 variables explained 75.21% of the total variance in algal-environment relation. The first axis of the ordination explained 55.9% of the total variance.

Results from CCA ordination of the most environmental variables and algal concentration show that WT, TP, pH and 8Hm were strongly correlated with the first CCA axis (0.5997, -0.6674, 0.4934 and -0.4559 respectively). TABLE Ⅶ shows the canonical structures of the first pairs of canonical variates. In Site 3#, the algal concentration represented Chl_a_1 was significantly affected by the environmental variables in one day before. According the result of CCA and domain knowledge, WT, TP, pH and 8Hm were selected to the input variables of Elman model of Site 3#.

TABLE VI. CANONICAL CORRELATION COEFFICIENTS IN SITE 3#

Axis	Axis1	Axis2	Axis3
Eigenvalues	1.6560	0.6234	0.2627
Species-environment correlations	0.7896	0.4561	0.2981
Cumulative (%)	55.90	60.68	75.21
Monte Carlo test: Eingenvalues-p	<.001	0.1775	0.0665

TABLE VII. CANONICAL STRUCTURES OF THE FIRST PAIR OF CANONICAL VARIATES IN SITE 3# (CC=0.7896(APPROX.F=1.66,P<0.001)

Environmental factors		Algal concentration	
variable	*loading*	*variable*	*loading*
WT(℃)	0.5997	chl_a_1	0.9232
pH	0.4934	chl_a_2	0.3380
DO (mgL^{-1})	0.2698	chl_a_3	0.1765
CODmn (mgL^{-1})	0.3290		
TN (mgL^{-1})	-0.1393		
TP (mgL^{-1})	-0.6674		
8Hm(cm)	-0.4005		
WQ	-0.4559		

B. Predictability of the Elman models

The Elman model was developed to simulate 1-day-ahead of Chl_a concentrations at three sites in Gonghu Bay of Lake Taihu. The architecture of the best Elman model for the Chl_a is shown in Fig.2. The Elman model is composed of one input layer with input variables selected by CCA, one hidden layer with optimized nodes and one output layer with one output variable. The parameters of Elman model which produced the "best results" for validation data set was were conducted according to Heyi Wang[9].

The developed Elman models accurately simulated the Chl_a concentrations at three sites in Gonghu Bay of Lake Taihu. The results are described in Fig.3. Using optimized input variables, the Chl_a concentrations prediction model accurately simulated the range of Chl_a concentrations at site 1(R^2=0.90; MRPE=19.42%), site 2(R^2=0.86; MRPE=17.61%) and site 3(R^2=0.92; MRPE=13.17%). The model simulated Chl_a concentrations with a good accuracy. The Elman model was able to simulate the Chl_a concentration with an accuracy of a degree or less (MRPE<20% and R^2>0.85). The result of Elman model shows that it is possible to predict algal dynamic variation in three sites in summer.

V. CONCLUSION

In this paper, using continuous daily measurements of environmental parameters at different sites, Elman models were created to imitate algal dynamic of Gonghu Bay during alga bloom. Based on CCA, the factors affecting the change of algal dynamic were selected to be input variables. In spite of largely unknown factors controlling transferring algal dynamic variation and the limited data set size, a relatively good correlation was observed between the measured and predicted values. In our study, CCA was first employed to examine interactions between the ecological factors that influence plankton communities in Gonghu Bay. The discussion shows that the Elman can be used to extract, recognize and predict related patterns of limnological time series. It is also stated that the input variables computed by CCA is acceptable. We suggest that the Elman can be as a powerful predictive alternative to traditional modelling techniques and the accuracy of the predictions is improved with increasing event and time resolution of training data. The successful application of Elman models to freshwater ecosystems may provide the opportunity of improving the efficiency of monitoring and management systems.

ACKNOWLEDGMENT

I wish to thank Bureau of Hydrology and Water Resources Monitoring, Taihu Basin Management Bureau, for providing data of the lake Taihu.

REFERENCES

[1] Asaeda, T., Bon, T.V., 1997. Modelling the effects of macrophytes on algal blooming in eutrophic shallow lakes. Ecol. Model. 104, 261-287.

[2] Webster, I.T., Sherman, B.S., Bormans, M., Jones, G., 2000. Management strategies for cyanobacterial blooms in an impounded lowland river. Regul. Rivers: Res. Manage. 16, 513-525.

[3] Maier, H.R., Burch, M.D., Bormans, M., 2001. Flow management strategies to control blooms of the cyanobacterium, Anabaena circinalis, in the River Murray at Morgan South Australia. Regul. Rivers: Res. Manage. 17, 637-650.

[4] Jeong, K.-S., Kim, D.-K., Whigham, P., Joo, G.-J., 2003a. Modelling Microcystis aeruginosa bloom dynamics in the Nakdong River by means of evolutionary computation and statistical approach. Ecol. Model. 161, 67-78.

[5] Recknagel, F., 1997. ANNA-Artificial Neural Network model for predicting species abundance and succession of blue-green algae. Hydrobiologia 349, 47-57.

[6] Maier, H.R., Dandy, G.C., 2001. Neural Network Based Modelling of Environmental Variables: A Systematic Approach. Mathematical and Computer Modeling 33, 669-682.

[7] Wei, B., Sugiura, N., Maekawa, T., 2001. Use of artificial neural network in the prediction of algal blooms. Water Research 35(8), 2022-2028.

[8] Walter, M., Recknagel, F., Carpenter, C., Bormans, M., 2001. Predicting eutrophication effects in the Burrinjuck Reservoir (Australia) by means of the deterministic model SALMO and the recurrent neural network model ANNA. Ecological Modeling 146 (1-3), 97-114.

[9] Heyi Wang, Yi Gao,Zhaoan Xu,Weidong Xu. 2011 International Conference on Remote Sensing, Environment and Transportation Engineering. 984-988

[10] Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Model. 178,389-397

[11] Jeong, K.-S., Kim, D.-K., Joo, G.-J., 2006. River phytoplankton prediction model by Artificial Neural Network: model performance and selection of input variables to predict time-series phytoplankton proliferations in a regulated river system. Ecol. Inform. 1, 235-245.

[12] Hansel-welch, N., Butler, M.G., Carlson, T.J., Hanson, M.A., 2003. Changes in macrophyte community structure in Lake Christina (Minnesota), a large shallow lake, following biomanipulation. Aquatic Botany 75, 323-337.

[13] Tan, X., Kong, F.X., Zeng, Q.F., Cao, H.S., Qian, S.Q., Zhang, M., 2009. Seasonal variation of Microcystis in Lake Taihu and its relationships with environmental factors. Journal of Environmental Sciences 21, 892-899.

[14] Qiao Shuliang, Jin man, Chen Guoping, Zou Shan. Calculation method and characteristics of wind-wave in lake. Hydro-Science and Engineering.1996(3),189-198.

[15] Marcel Bottema, Gerbrant Ph. van Vledder. A ten-year data set for fetch- and depth-limited wave growth. Coastal Engineering 56 (2009) ,703-725

[16] Hu,W.,Pu,P.,Qin,B.,1998.A three-dimensional numerical simulation on the dynamics in Taihu Lake, China(I): the water level and the current during the 9711 typhoon process.J.LakeSci.4,17-25(in Chinese with English abstract).

[17] Vanderpoorten, A., Palm, R., 1998. Canonical variables of aquatic bryophyte combinations for predicting water trophic level. Hydrobiologia 386, 85-93.

[18] Recknagel, F., French, M., Harkonen, P., Yabunaka, K.-I., 1997. Artificial neural network approach for modeling and prediction of algal blooms. Ecol. Model. 96, 11-28.

[19] Tarassenko, L., 1998. A Guide to Neural Computing Applications. Arnold Publishers, London.

[20] Hecht-Nielsen, R., 1987. Kolmogorov's mapping neural network existence theorem. Proceedings of 1st IEEE International Joint Conference of Neural Networks. Institute of Electrical and Electronics Engineers, New York, NY.

TABLE VIII. SOME STATISTICS ON THE MEASURED DATA

	WT(℃)	pH	DO (mgL^{-1})	COD_{MN} (mgL^{-1})	TN (mgL^{-1})	TP (mgL^{-1})	Chl_a(mgm^{-3})	8Hm(cm)
Site 1# (120°13'51.0"E，31°23'51.8"N)								
Mean	26.54	8.23	6.97	3.79	1.61	0.06	14.83	8.96
Min	18.50	7.51	4.69	2.18	0.65	0.02	2.50	1.76
Max	33.10	9.33	10.94	13.20	4.55	0.15	73.40	28.29
Site 2# (120°22'17.8"E，31°26'46.8"N)								
Mean	26.66	8.33	7.40	4.03	1.62	0.08	27.68	9.00
Min	17.9	7.21	5.22	2.44	0.72	0.03	2.80	1.76
Max	33.5	9.18	10.8	9.98	3.74	0.40	77.50	28.29
Site 3# (120°22'32.5"E，31°22'50.6"N)								
Mean	26.69	8.59	7.74	3.08	1.03	0.02	4.67	8.92
Min	18.30	7.50	4.58	2.18	0.29	0.01	1.00	1.76
Max	33.40	9.51	12.84	4.36	3.12	0.07	24.10	28.29

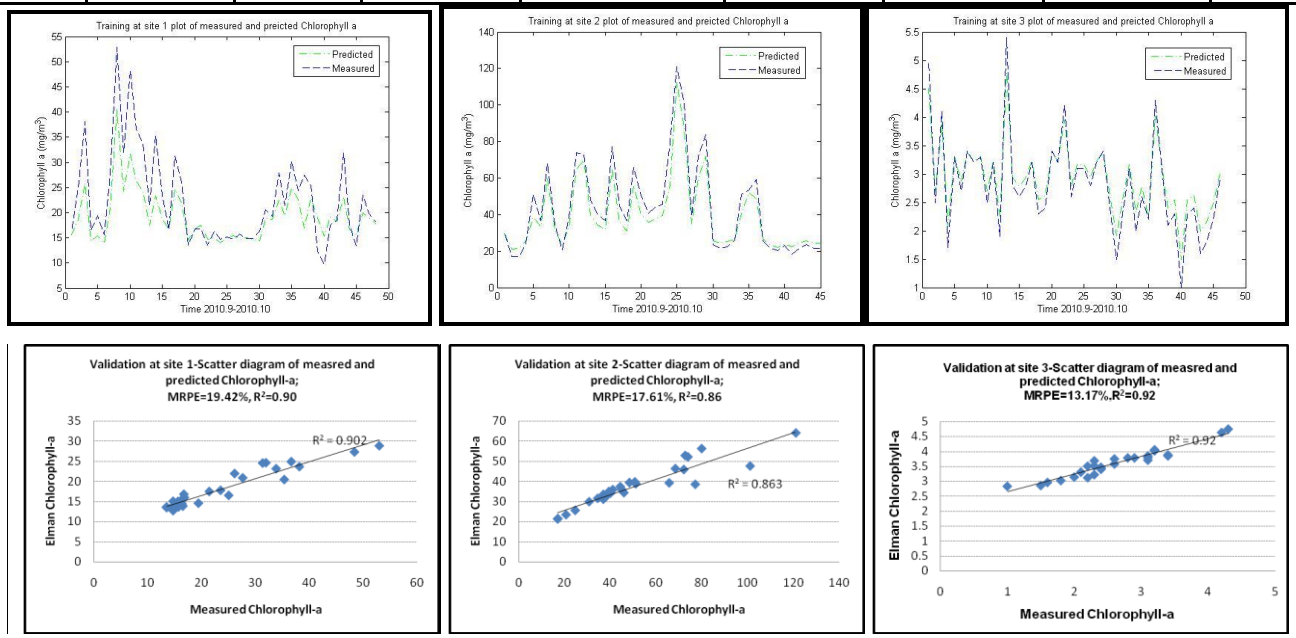

Fig.3 Measured and predicted Chl_a concentrations for training and validation tests of three sites