J. Agr. Sci. Tech. (2016l. 18: 975-983

Water Quality Analysis and Prediction Using Hybrid Time Series and Neural Network Models

L. Zhang[1], G. X. Zhang[1]*, and R. R. Li[1]

ABSTRACT

Chagan Lake serves as an important ecological barrier in western Jilin. Accurate water quality series predictions for Chagan Lake are essential to the maintenance of water environment security. In the present study, a hybrid AutoRegressive Integrated Moving Average (ARIMA) and Radial Basis Function Neural Network (RBFNN) model is used to predict and examine the water quality [Total Nitrogen (TN), and Total Phosphorus (TP)] of Chagan Lake. The results reveal the following: (1) TN concentrations in Chagan Lake increased slightly from 2006 to 2011, though yearly variations in TP were not significant. The TN and TP levels were mainly classified as Grades IV and V, (2) The hybrid ARIMA and RBFNN model's *RMSE* values for the observed and predicted data were 0.139 and 0.036 mg L^{-1} for TN and TP, respectively, which indicated that the hybrid model describes TN and TP variations more comprehensively and accurately than single ARIMA and RBFNN model. The results serve as a theoretical basis for ecological and environmental monitoring of Chagan Lake and may help guide irrigation district and water project construction planning for western Jilin Province.

Keywords: ARIMA model, Chagan Lake, RBFNN model, Total N, Total P.

INTRODUCTION

Accurate water quality predictions guide water quality management decisions, aquaculture water plans, and water quality incident strategies. Studies focusing on intensive aquaculture water quality prediction methods are thus of critical theoretical value and practical significance (Xu and Liu, 2013). For water quality prediction models, frequently used methods include regression models (Abaurrea *et al*., 2011), grey water quality models (Karmakar and Mujumdar, 2006), time series models (Ahmad *et al*., 2001), and Artificial Neural Network (ANN) models (May *et al*., 2008). One of the most prominent and widely used time series models is the Box-Jenkins modeling approach, commonly known as the AutoRegressive Integrated Moving Average (ARIMA) (Box *et al*., 1994). ARIMA models are flexible in that they can depict several different time series, i.e., pure AutoRegressive (AR), pure Moving Average (MA) and combined AR and MA (ARMA) series, though they are limited by their pre-assumed linear form (Parviz *et al.*, 2010). Because a linear correlation structure is assumed between the time series values, ARIMA models cannot address nonlinear relationships. Linear model approximations for complex real-world problems are not always satisfactory (Zhang, 2003). Neural Networks (NNs) can identify complex nonlinear relationships between input and output datasets without requiring information on the nature of the phenomena and without making underlying assumptions regarding linearity or normality. However, a neural network model alone cannot address both linear and nonlinear patterns equally

[1] Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, 130102, People's Republic of China.
* Corresponding author; e-mail: zhgx@iga.ac.cn

[DOR: 20.1001.1.16807073.2016.18.4.21.5]

well (Al-Alawi *et al*., 2008). Therefore, by combining ARIMA and ANN models, the complex autocorrelation structures in data can be modeled more accurately.

Chagan Lake, a typical shallow soda-saline lake in the semi-arid region of southwestern Songnen Plain, Northeast China, serves as an important ecological barrier in western Jilin Province and as the most important fishery base in Jilin Province. Shen and Zhang (2009) concluded that enclosure development and construction in Chagan Lake and the implementation of water conservancy projects resulted in serious ecological degradation in the lake. Several studies have also identified serious TN, TP and (Permanganate index) COD_{Mn} pollution levels in the lake (Dai and Tian, 2011). Chagan Lake is connected to the Songhua River by irrigation channels, through which the lake receives a large amount of water from the Second Songhua River and from the Qianguo Irrigation Area agricultural drainage, which is also sourced from the river (Zhu *et al*., 2012). With the development of the saline-alkali land, large volumes of farmland drainage with high concentrations of TN, TP, and salt will flow into Chagan Lake and will inevitably affect the water quality and ecological security of Chagan Lake. Effective water quality predictions for Chagan Lake following the development of the Songyuan irrigation district must be conducted to ensure water safety for the purposes of sustainable development and public health.

This study aimed to execute the following tasks: (1) Analysis of the water quality fluctuations in Chagan Lake in recent years by focusing on TN and TP levels; (2) Develop a hybrid ARIMA and RBFNN model to predict water quality time series data; and (3) Assess the performance of these models by comparing observed and predicted data, thereby evaluating the predictive performance of the hybrid ARIMA and RBFNN model relative to the ARIMA model.

MATERIALS AND METHODS

Study Area and Dataset

Chagan Lake, located in the southwest area of the Songnen Plain, Northeast China (124° 03′ 28″-124° 30′ 59″ E, 45° 05′ 42″- 45° 25′ 50″ N; Figure 1), is the tenth largest lake in China. The lake covers a mean surface area of 372 km^2, has a mean depth of 1.52 m, and

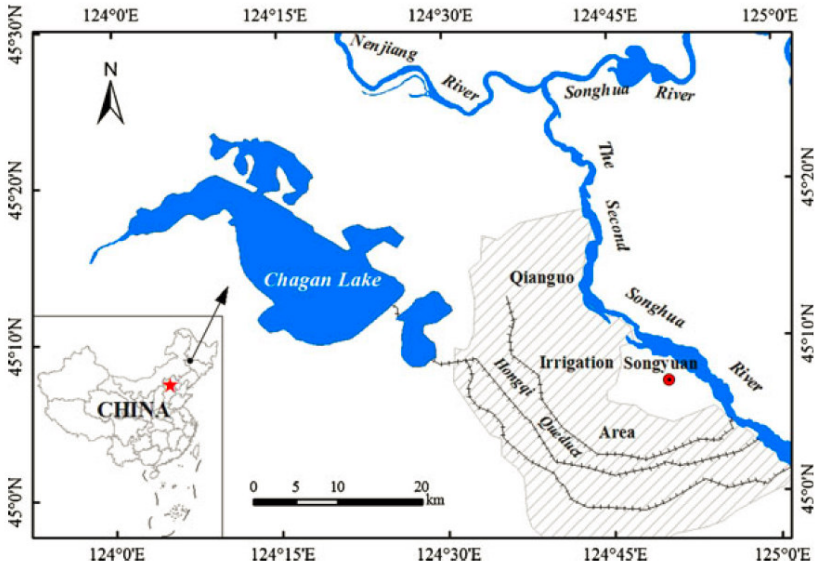

Figure 1. Map of Chagan Lake (Zhu *et al.*, 2012)

[DOR: 20.1001.1.16807073.2016.18.4.21.5]

features a full storage capacity of 5.98×10^8 m^3. The regional climate is categorized as a mid-temperate zone. The annual mean air temperature for the area is 4.5ºC, and the local freezing period lasts from October to the following May. Chagan Lake is located in the semi-arid area of western Jilin Province. Hence, annual runoff levels are low, and average runoff depths of 5-10 mm have been recorded for several years (Zhu *et al.*, 2012).

The Second Songhua, Huolin, Tao'er and Nenjiang Rivers are tributaries of Chagan Lake, and natural precipitation and ground water serve as auxiliary water supplies for the lake. The Second Songhua River (primarily composed of farmland drainage from the Qianguo irrigation district) first flows into Lake Xinmiao through a canal and then into Chagan Lake. The river serves as the main source of water for the lakes in the region (Duan *et al*., 2008).

In this paper, the water quality parameters TN and TP are examined. Monthly data for these parameters for the period 2006-2011 were used for the analysis. Water quality data were obtained from "The Second Songhua River Diversion Project Record" (A and En, 2003) from the Hydrology Bureau of Jilin Province, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences. The statistical properties of the water quality time series data and the environmental quality standards for surface water of the People's Republic of China (GB3838-2002 of China P. R. 2002) are presented in Table 1.

Hybrid ARIMA and RBFNN Models

It may be reasonable to consider a time series (y_t) to be composed of a Linear autocorrelation structure (L_t) and a Nonlinear component (N_t). That is,

$$y_t=L_t+N_t \quad (1)$$

There were three steps to predict the water quality series by hybrid ARIMA and RBFNN models.

1) The ARIMA model (Box *et al*., 1994) was used to predict y_t, and let $\hat{L}_t$ denote the prediction results. The e_t was the residuals between the ARIMA model and series.

$$e_t = y_t - \hat{L}_t \quad (2)$$

2) The e_t was considered as the input of RBFNN model (Moody and Darken, 1989), then the RBFNN model could be expressed as follows:

$$e_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-n}) + \varepsilon_t \quad (3)$$

Where, f is a nonlinear function determined by the neural network and $\_\varepsilon_t$ is the random error.

The output results of RBFNN was defined as $\hat{N}_t$.

3) The two models were combined for forecast, and the prediction results from hybrid ARIMA RBFNN models were expressed as:

$$\hat{y}_t = \hat{L}_t + \hat{N}_t \quad (4)$$

So, the predicted water quality results generated through the hybrid ARIMA-RBFNN model were obtained through a combination of the linear prediction by ARIMA and the ARIMA model residuals with nonlinear characteristic predicted by RBFNN model prediction results (Zhang, 2003). This procedure is illustrated in Figure 2.

Table 1. Statistical properties of the water quality parameters and the environmental quality standards for surface waters, China (GB3838-2002).

Parameters	Min value	Max value	Mean	Std dev	Environmental quality standard					
						I	II	III	IV	V
TN (mg L^{-1})	0.24	2.81	1.29	0.58	≤	0.2	0.5	1.0	1.5	2.0
TP (mg L^{-1})	0.02	0.51	0.10	0.08	≤	0.01	0.025	0.05	0.1	0.2

[DOR: 20.1001.1.16807073.2016.18.4.21.5]

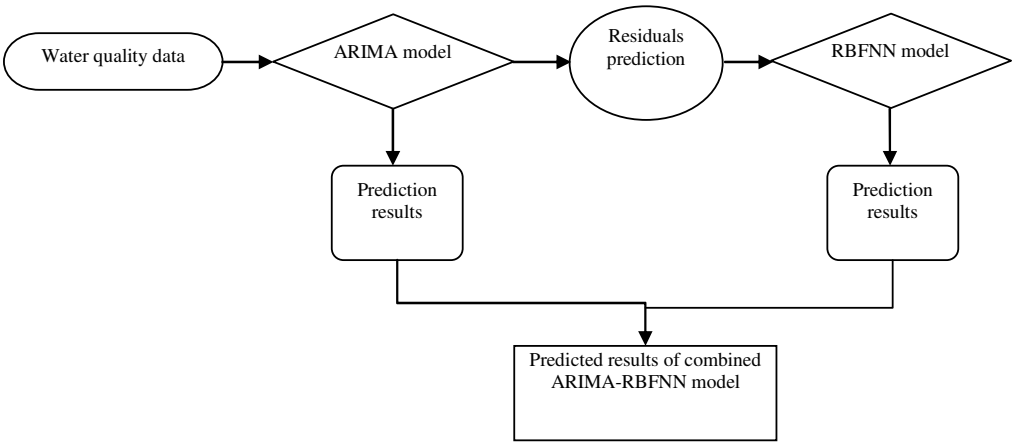

Figure 2. Schematic diagram of the ARIMA - RBFNN model.

RESULTS AND DISCUSSION

Variation Characteristics of the Water Quality Time Series for Chagan Lake

The tendency of TN and TP year after year and month by month from 2006 to 2011 is shown in Figure 3.

The TN, TP, N/P levels in Chagan Lake showed similar variation trend, with increasing from 2006 to 2008, then decreasing slightly from 2008 to 2011. The yearly variations in TN and TP, classified as between Grades IV and V. The yearly N/P variations were relatively stable in Chagan Lake, typically exceeding 16:1, suggesting phosphorus was the limiting nutrient.

The monthly TN levels in Chagan Lake showed an overall decreasing trend, classified as between Grades IV and V, though this was not the case during the months of September and November. The maximum values were recorded during the coldest period of the year, and levels showed a short-term increase during the spring months in association with the influx of alkali due to irrigation. The monthly TP levels in Chagan Lake increased in the summer and autumn months, but decreased in the spring and winter. The peak values were not synchronized with TN during the peak chemical fertilization period between May and June but were synchronized with the presence of large channel water volumes between August and October. The monthly N/P variations in Chagan Lake significantly declined throughout the year until late fall. An average N/P value of 10.43 was observed for June to October, creating suitable conditions for aquatic plant growth, especially algae growth.

Development of Hybrid ARIMA and RBFNN Models for Chagan Lake

ARIMA Modeling

According to the ACF and PACF diagrams for TN (Figure 4), the following preliminary ARIMA model parameters were identified: p= 1-2, q=1, and d= 1 (Cryer and Chan, 2008). The matching test showed that p= 1, d= 1, and q= 1, forming the ARIMA model for TN prediction in Chagan Lake [ARIMA (1,1,1)]. For TP, the following preliminary ARIMA model parameters were identified: p= 1-4, q= 1, and d= 1. The matching test showed that p= 2, d= 1, and q= 1. The ARIMA model for TP prediction in Chagan Lake was identified as ARIMA (2,1,1). Using the 2006-2010 water quality series for Chagan Lake as a training sample, the 2011 water quality parameters were predicted.

[DOR: 20.1001.1.16807073.2016.18.4.21.5]

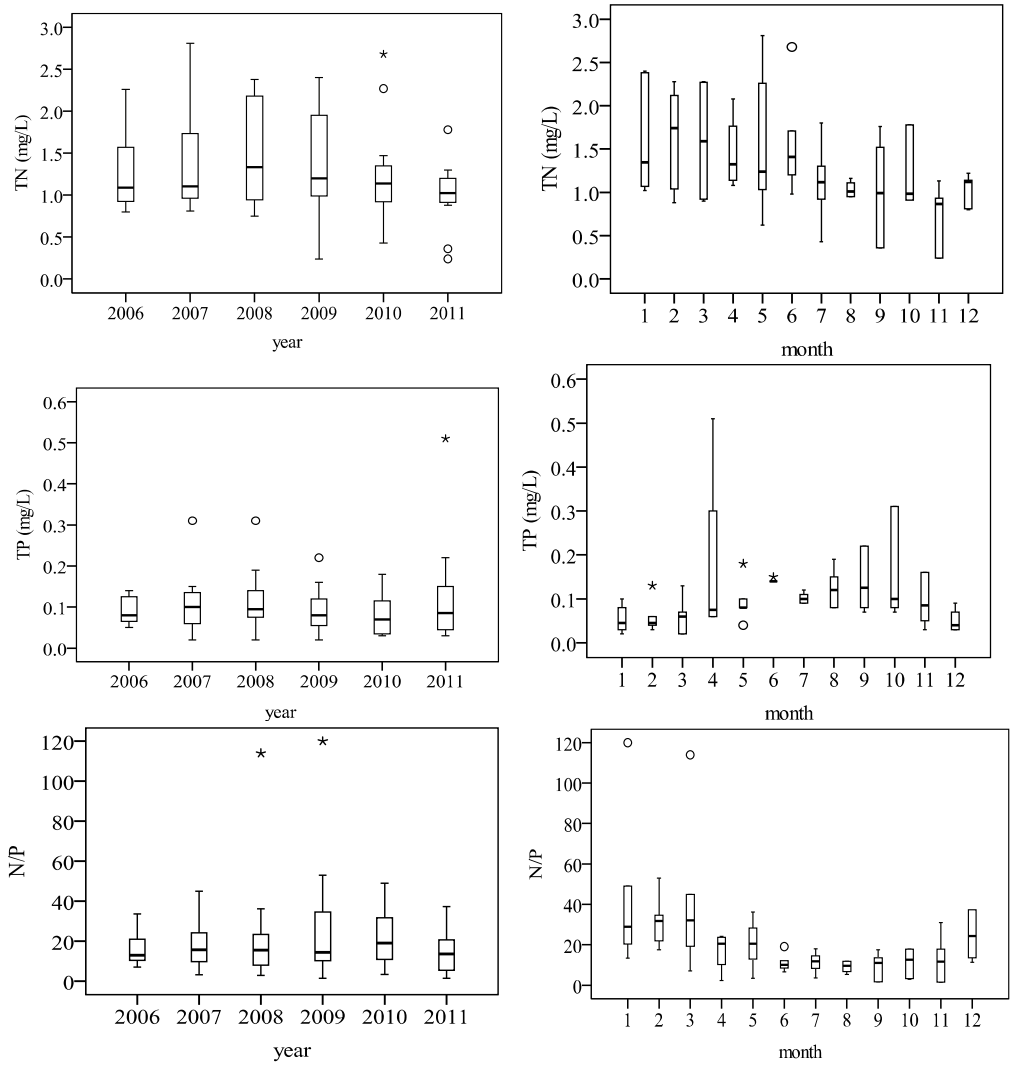

Figure 3. Temporal variations in TN, TP and N/P levels in Chagan Lake.

RBFNN Prediction

The prediction residual sequence for the water quality factors was obtained based on the ARIMA (1,1,1) prediction results and observations. The residual sequence was then used as the RBFNN input cell. The width of training σ was 0.6, and the number of nodes in hidden layers was 2.

The Hybrid ARIMA-RBFNN Model

The prediction residual sequence for the water quality factors was obtained based on the ARIMA (1,1,1) prediction results and observations. The residual sequence was then used as the RBFNN input cell. Finally, the hybrid ARIMA-RBFNN model was determined through the linear superposition of the ARIMA-predicted values and the RBFNN-derived ARIMA residual prediction values. The ARIMA, RBFNN and hybrid ARIMA-RBFNN model prediction results are shown in Figure 5.

The RBFNN showed extremely bad prediction results for TP. Although it identified the variation trend of TN, the prediction results was also not satisfied. Though the prediction accuracy of the ARIMA was not high, the model identified

[DOR: 20.1001.1.16807073.2016.18.4.21.5]

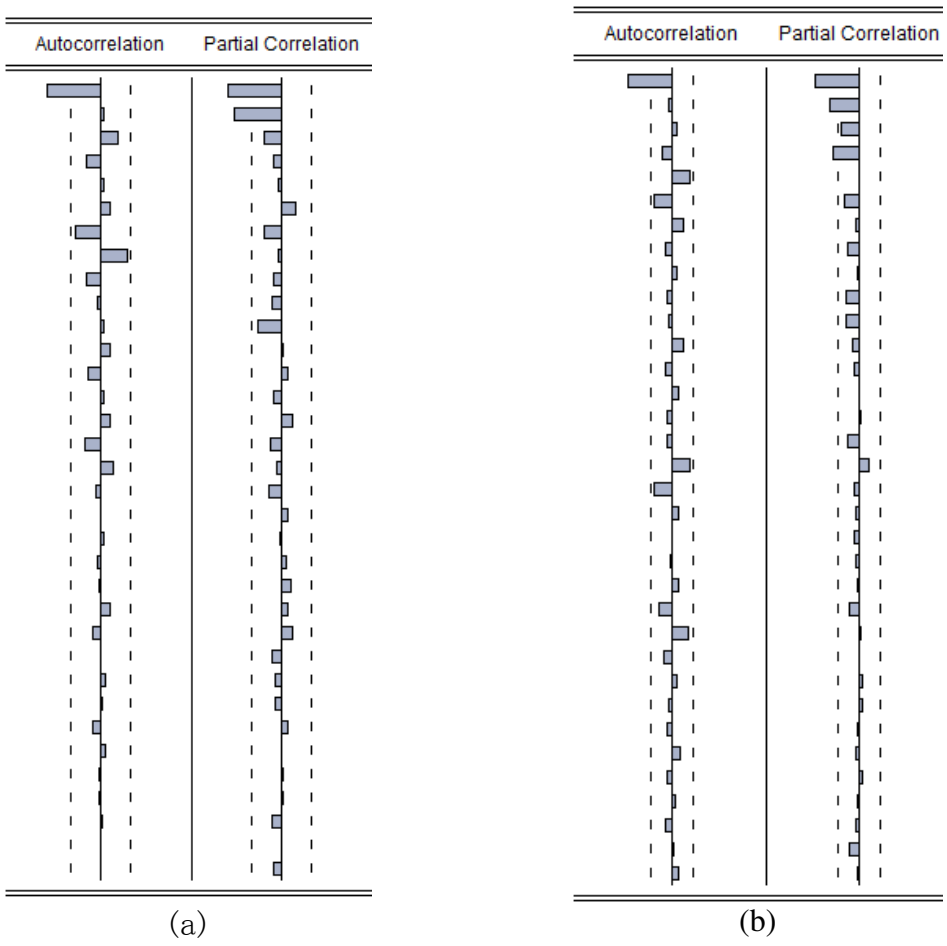

Figure 4. ACF and PACF diagrams for TN (a) and TP (b).

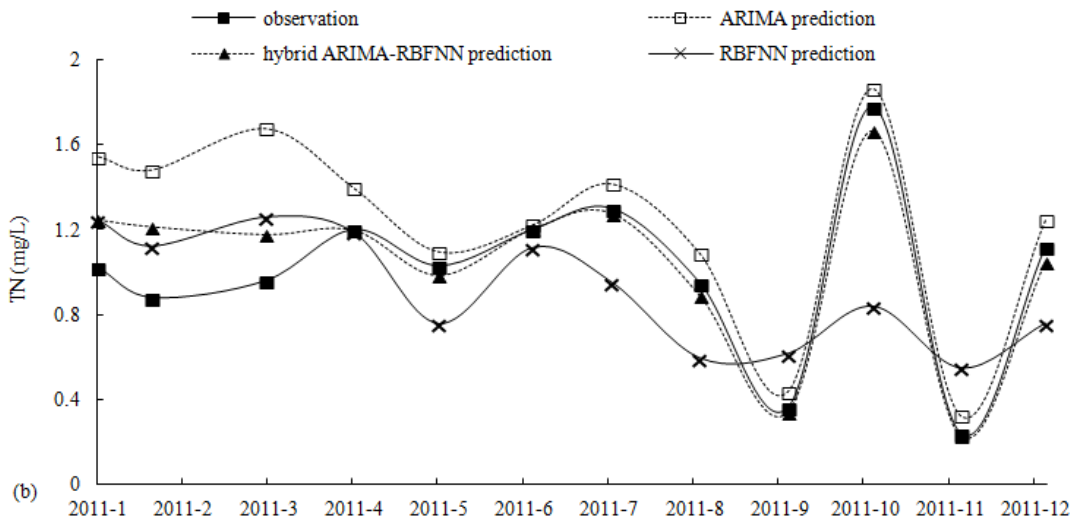

Figure 5. Comparison of the observed data and prediction results generated using the ARIMA, RBFNN and hybrid ARIMA-RBFNN models: (a) TP, (b) TN.

[DOR: 20.1001.1.16807073.2016.18.4.21.5]

TN and TP variation trends i.e., linear components of the TN and TP values. The TN prediction error of the ARIMA model was overwhelmingly larger for January through April than for the other months, whereas the maximum TP prediction error occurred in April. The elemental analysis results show that this is attributable to significant TN and TP concentration variations occurring in January through April during the period 2006-2010. Additionally, because TN pollution sources vary during the winter months, periodic variations during this period are not obvious. TP fluctuations found in April may be attributable to non-point source pollution released through snowmelt. Although the hybrid model produced superior predictions to those of the ARIMA, significant errors appeared between January and April. The time series analysis only considered the influence of historical data on future trends and did not directly consider the effects of various factors on the time series. Thus, the time series analysis largely depends on historical data. If environmental conditions change significantly, the influence of various factors on the time series will change, and the time series model will make accurate predictions. Time series analyses, therefore, cannot respond to the effects of sudden disturbances.

Comparison of Model Performance

In comparing the predicted and observed 2011 data from the ARIMA, RBFNN and hybrid ARIMA-RBFNN model, accuracy measures were employed to determine the best model by using the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) (Makridakis *et al*., 1986).

Table 2 presents the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) results for the observed and predicted TN and TP data in 2011 from the ARIMA, RBFNN, and hybrid ARIMA-RBFNN models. The RBFNN model's RMSE values for the observed and predicted data were 0.378 and 0.141 mg L^{-1} for TN and TP, respectively. The ARIMA model's *RMSE* values for the observed and predicted data were 0.324 and 0.104 mg L^{-1} for TN and TP, respectively. And hybrid model's *RMSE* values for the observed and predicted data were 0.139 and 0.036 mg L^{-1} for TN and TP, respectively. So, the hybrid model shows the best prediction results, and RBFNN the worst. From the *MAPE* results, furthermore, the observed and ARIMA-predicted error statistics produced *MAPE* values of 18.194 and 27.299% for TN and TP, respectively. The hybrid model generated *MAPE* values of 7.017 and 14.528% for TN and TP, respectively. The *RMSE* and *MAPE* values for the hybrid model were, therefore, lower than those of the ARIMA model and RBFNN model, suggesting that the predictive capacity of the hybrid model is superior to that of the ARIMA model and RBFNN model. The hybrid model effectively simulated TN and TP time series for Chagan Lake.

CONCLUSIONS

The periodic, nonlinear and random analyses of water quality data for Chagan Lake shows that the concentration of TN and TP are Grades IV and V. Compared to the ARIMA model, the proposed hybrid

Table 2. Comparison of the ARIMA, RBFNN and hybrid ARIMA-RBFNN models: *RMSE* and *MAPE* values.

Parameters	*RMSE*			*MAPE*		
	ARIMA	RBFNN	Hybrid ARIMA-RBFNN	ARIMA	RBFNN	Hybrid ARIMA-RBFNN
TN (mg L^{-1})	0.324	0.378	0.139	18.194	34.633	7.017
TP (mg L^{-1})	0.104	0.141	0.036	27.299	126.957	14.528

[DOR: 20.1001.1.16807073.2016.18.4.21.5]

ARIMA-RBFNN prediction model described TN and TP variations more comprehensively and accurately, producing TN and TP *RMSE* values that were 0.139 and 0.036 mg L^{-1}, respectively, than those generated using the ARIMA and *MAPE* values for TN and TP that were 7.017 and 14.528%, respectively. The hybrid model improved the prediction precision. The proposed hybrid model can therefore be used to predict TN and TP time series for Chagan Lake. These prediction results will serve as a theoretical basis for the ecological and environmental management of Chagan Lake and as a guide for the irrigation district and water project construction planning in western Jilin Province.

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant No.41371108 and No.41301086), the Scientific Research Project of Public Welfare Industry of the Ministry of Water Resources, China (No. 201401014).

REFERENCES

1. A, R. H. and En, H. 2003. The Second Songhua River Diversion Project Record of Qian Gorlos Mongol Autonomous County. Liaoning Minorities Press, Liaoning. (in Chinese)
2. Abaurrea, J., Asín, J., Cebrián, A. C. and García-Vera, M. A. 2011. Trend Analysis of Water Quality Series Based on Regression Models with Correlated Errors. *J. Hydrol.*, **400 (3)**: 341–352.
3. Ahmad, S., Khan, I. H. and Parida, B. P. 2001. Performance of Stochastic Approaches for Forecasting River Water Quality. *Water Resour.*, **35**: 4261-4266.
4. Al-Alawi, S. M., Abdul-Wahab, S. A. and Bakheit, C. S. 2008. Combining Principal Component Regression and Artificial Neural Networks for More Accurate Predictions of Ground-Level Ozone. *Environ. Model. Softw.*, **23(4)**: 396-403.
5. Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. 1994. *Time Series Analysis, Forecasting and Control.* Prentice Hall, Englewood Cliffs, NJ.
6. Cryer, J. D., Chan, K. S. 2008. Time Series Analysis: With Applications in R, second ed. Springer, New York.
7. Dai, X. J. and Tian, W. 2011. Analysis and Countermeasures on Water Pollution of Lake Chagan. *J. Arid Land Resour. Environ.*, **25(8)**: 179-184. (in Chinese)
8. Duan, H., Zhang, Y., Zhang, B., Song, K., Wang Z., Liu, D. and Li, F. 2008. Estimation of Chlorophyll-a Concentration and Trophic States for Inland Lakes in Northeast China from Landsat TM Data and Field Spectral Measurements. *Int. J. Remote Sens.*, **29**:767–786.
9. GB3838-2002 of China P. R. 2002. *Environmental Quality Standard for Surface Water. China: Environmental Science*. Ministry of Environmental Protection of the People's Republic of China, China.
10. Karmakar, S. and Mujumdar, P. P. 2006. Grey Fuzzy Optimization Model for Water Quality Management of a River System. *Adv. Water Resour.*, **29**: 1088–1105.
11. Makridakis, S., Wheelwright, S. C. and Megee, V. E. 1986. *Forecasting: Method and Application*. Second Edition, Wiley, New York.
12. May, R. J., Dandy, G. C., Maier, H. R., Nixon, J. B. 2008. Application of Partial Mutual Information Variable Selection to ANN Forecasting of Water Quality in Water Distribution Systems. *Environ. Modell. Softw.*, **23(10–11)**: 1289-1299.
13. Moody, J. and Darken, C. 1989. Fast Learning in Networks of Locally Tuned Processing Units. *Neural Comput.*, **1**: 281–294.
14. Parviz, L., Kholghi, M. and Hoorfar, A. 2010. A Comparison of the Efficiency of Parameter Estimation Methods in the Context of Stream Flow Forecasting. *J. Agr. Sci. Tech.,* **12(1)**: 47-60.
15. Shen, J. L. and Zhang, J. L. 2009. Ecosystem Research of Chaganhu Conservation Zone Based on Remote Sensing Data and Object-oriented Classification. *J. Earth Sci. Environ.*, **31(2)**: 212-215. (in Chinese)
16. Xu, L. Q. and Liu, S. Y. 2013. Study of Short-term Water Quality Prediction Model Based on Wavelet Neural Network. *Math. Comput. Model.*, **58(3-4)**: 807–813.

[DOR: 20.1001.1.16807073.2016.18.4.21.5]

17. Zhang, G. P. 2003. Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. *Neurocomput.*, **50**: 159-175.
18. Zhu, L. L., Yan, B. X.,·Wang, L. X. and Pan, X. 2012. Mercury Concentration in the Muscle of Seven Fish Species from Chagan Lake, Northeast China. *Environ.Monit. Assess.*, **184**: 1299-1310.

تجزیه و پیش بینی کیفیت آب با استفاده از مدل هیبرید سری زمانی و مدل شبکه عصبی

ل. ژانگ، ج. ز. ژانگ، و ر. ر. لی

چکیده

دریاچه چاگان در غرب منطقه جیلین به عنوان یک مانع اکولوژیکی مهم عمل می کند و برای حفظ امنیت محیط آبی این دریاچه پیش بینی درست سری زمانی کیفیت آب ضرورت دارد. در پژوهش حاضر، برای پیش بینی و آزمون کیفیت آب [نیتروژن کل (TN) و فسفر کل(TP)] دریاچه چاگان از میانگین های متحرک تلفیقی خود رگرسیونی (AutoRegressive Integrated Moving Average: ARIMA) و مدل شبکه عصبی تابع مبتنی بر شعاع (Radial Basis Function Neural Network (RBFNN) استفاده شد. نتایج حاکی از موارد زیر بود: (۱) غلظت های TN در دریاچه چاگان از سال ۲۰۰۶ تا ۲۰۱۱ اندکی روندی افزایشی داشت، هرچند که تغییرات سالانه در TPمعنی دار نبودند. همچنین، مقادیر TNو TPدر رده کلاس چهار و پنج طبقه بندی شدند، (۲) مقدار *RMSE*در مورد داده های مشاهده شده و پیش بینی شده TNو TPبه دست آمده ازهیبرید ARIMA وRBFNN به ترتیب برابر ۰/۱۳۹ و۰/۰۳۶ میلی گرم در لیتر بود که حاکی از آن بود که مدل هیبریدی مزبور تغییرات TNو TPرا درست تر و کامل تر از زمانی که ARIMA یا مدل RBFNN به تنهایی استفاده شوند توضیح می دهد. این نتایج به عنوان پایه نظری برای پایش اکولوژیکی و محیط زیستی دریاچه چاگان عمل می کند و می تواند به مسولان آبیاری آن ناحیه و در برنامه ریزی ساخت پروژه های آب برای استان جیلین غربی کمک کند.

[DOR: 20.1001.1.16807073.2016.18.4.21.5]