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

L. Zhang¹, G. X. Zhang¹*, and R. R. Li¹

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.

¹Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun, 130102, People’s Republic of China.
* Corresponding author; e-mail: zhgx@iga.ac.cn
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\textsubscript{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 Qianhuo 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\textsuperscript{2}, has a mean depth of 1.52 m, and

![Figure 1. Map of Chagan Lake (Zhu et al., 2012)](image-url)
Water Quality Prediction by ARIMA-RBFNN Model

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

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Min value</th>
<th>Max value</th>
<th>Mean</th>
<th>Std dev</th>
<th>Environmental quality standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>TN (mg L⁻¹)</td>
<td>0.24</td>
<td>2.81</td>
<td>1.29</td>
<td>0.58</td>
<td>≤ 0.2 0.5 1.0 1.5 2.0</td>
</tr>
<tr>
<td>TP (mg L⁻¹)</td>
<td>0.02</td>
<td>0.51</td>
<td>0.10</td>
<td>0.08</td>
<td>≤ 0.01 0.025 0.05 0.1 0.2</td>
</tr>
</tbody>
</table>

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,
\[
y_t = L_t + N_t
\] (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.
\[
e_t = y_t - \hat{L}_t
\] (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:
\[
e_t = f(e_{t-1}, e_{t-2}, \ldots, e_{t-n}) + \epsilon_t
\] (3)

Where, \( f \) is a nonlinear function determined by the neural network and \( \epsilon_t \) 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:
\[
\hat{y}_t = \hat{L}_t + \hat{N}_t
\] (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.
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.
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 $\sigma$ 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...
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.
Table 2. Comparison of the ARIMA, RBFNN and hybrid ARIMA-RBFNN models: RMSE and MAPE values.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>RMSE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ARIMA</td>
<td>RBFNN</td>
</tr>
<tr>
<td>TN (mg L⁻¹)</td>
<td>0.324</td>
<td>0.378</td>
</tr>
<tr>
<td>TP (mg L⁻¹)</td>
<td>0.104</td>
<td>0.141</td>
</tr>
</tbody>
</table>

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 model showed superior results. The hybrid model effectively simulated TN and TP time series for Chagan Lake.
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

Water Quality Prediction by ARIMA-RBFNN Model


چکیده

دریابه چاگان در غرب منطقه چلین به عنوان یک منبع اکولوژیکی مهم عمل می‌کند و برای حفظ محیط آبی این دریابه پیش‌بینی درست سری زمانی کیفیت آب ضرورت دارد. در پژوهش حاضر، برای پیش‌بینی و آزمون کیفیت آب [تی‌ام‌اس از (TP) و فسفر کل (TN)] در دریابه چاگان از مدل‌هایی مثل تلقیه خود رگرسیونی (ARIMA: Average AutoRegressive Integrated Moving Average) و الگوی شبکه عصبی تابع مینیمیمیت بر شعاع (RBFNN: Radial Basis Function Neural Network) استفاده شد. نتایج حاصل از مورد زیر بود: (1) تایید های TN کمتر در دو سال 2006 و 2011، 2011 از ونیزی در تاریک بود، هرچند که تغییرات سالانه در مرغی دار نیویورک، همچنین، مقدار RNSE در مورد داده‌های مشاهده شده و پیش‌بینی شده RNSE مقدار RNSE به ترتیب بهره 139/0 و 139/0 تولید گرم در لیتر بود که حاکی از آن بود که رابطه RNSE با ARIMA مدل هیبریدی مزبور تغییرات RNSE در دو سال و کاملاً از زمانی که تا مدلRNSE به تهیه این استفاده شوند توضیح می‌دهد. این نتایج به عنوان یک نظیر برای پایان برنامه ریزی ساخت پروژه‌های آب برای استان چلین غربی کمک کند.