Prediction of total nitrogen in Gucheng Lake using artificial neural networks combined with factor correlation analysis

X. Wen*, G. Fang*

* College of Water Conservancy and Hydropower Engineering, Hohai University, 1 Xikang Road, Nanjing City 210098, China (E-mail: *njwenxin@163.com)*

Abstract
Gucheng Lake has been suffering from eutrophication due to increased pollution and nutrient loads discharged into the watershed. Based on artificial neural networks (ANNs) and a 4-year record of water quality data, this study proposes an early-warning model for eutrophication aiming to predict the concentration of total nitrogen (TN) of Gucheng Lake with a lead time of one week. To develop such data-driven models efficiently, a comprehensive sampling strategy is adopted to ensure that most relevant predictors for TN are retained. Factor correlation analysis is then employed to further eliminate noisy predictors. The preferable selecting ranges of correlation coefficient values are proven to be [-1, -0.5] and [0.5, 1]. As a result, 6 input variables are filtered from 75 potential input variables to develop the TN prediction models. The prediction models can achieve high performance. The validation results of TN showed that the correlation coefficient of 0.9915 and the RMSE of 0.0684, which have demonstrated the potential of ANN models to predict TN conditions at Gucheng Lake.

Keywords
artificial neural network; eutrophication; factor correlation analysis; prediction; water quality

Introduction

Eutrophication, the process of increasing organic enrichment of an ecosystem (Nixon, 1995) and generally attributed to the excessive nutrient load of nitrogen (Carpenter et al, 1998), has shown adverse impact on the eco-environmental system and water quality in Gucheng Lake. The main goal of this study is to predict the concentration of TN of Gucheng Lake one week in advance, using ANNs combined with factor correlation analysis. The sampling strategy of the study area is first developed based on the nutrient sources analysis. Afterwards, the modeling methods are described including the ANN structure, back propagation algorithm, parameter selection, and performance measures. In addition, factor correlation analysis used to filter input variables is also introduced. Then, the prediction results of the ANN models are presented, followed by the discussions and conclusions in the final section.

MATERIAL AND METHODS

Research area

Gucheng Lake is a shallow grass-type lake situated in the lower reach of the Yangtze River at East China (East: 118°41′-119°12′, North: 31°13′-31°26′), as shown in Figure 1. At its normal storage level, the lake covers 26.35km^2 with an average depth of 1.57 m and storage of 7.66×10^7 m^3. The hydrological retention time is about 3 months.

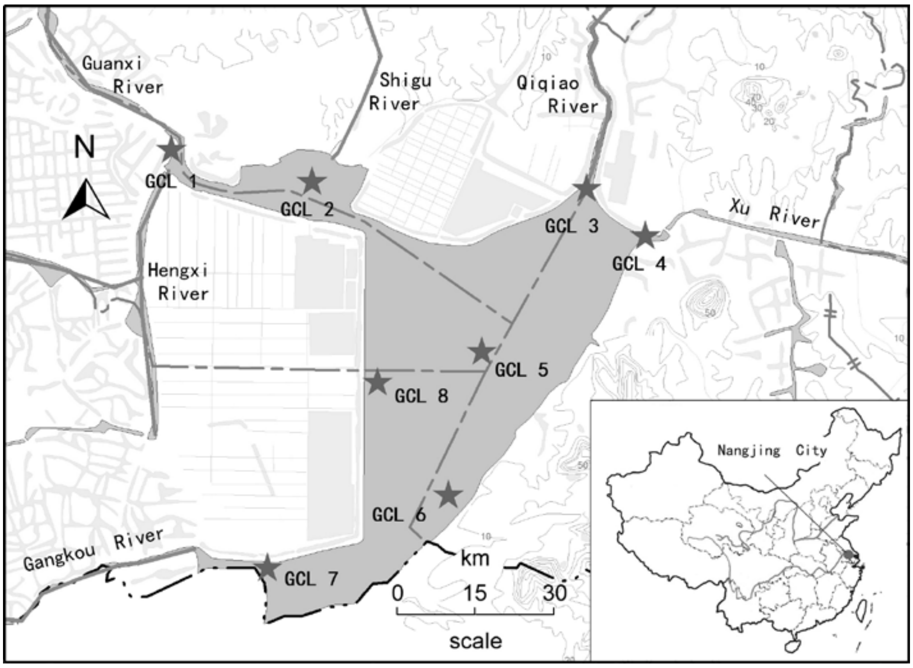

Figure 1. Water quality monitoring stations at Gucheng Lake basin

Methods

Artificial Neural Network is a data-driven model that tries to simulate the structural and functional aspects of biological neural networks. Multilayer perceptron (MLP) is composed of one input layer, introducing the basic information to the model, one hidden layer, performing the computation and simulation, and one output layer, presenting the results(Melesse 2008). These layers comprise of an interconnected group of neurons that processes information using a connectionist approach to computation.

At Site GCL5 (lake center), TN is chosen as predicting variables with a lead time of 1 week. Factor correlation analysis is introduced to investigate the statistical relationship between every potential variable and the target output, and then, to eliminate those noisy data by setting selecting range of correlation coefficients of each potential variable. Specifically, Pearson product-moment correlation coefficient (r) and Spearman's rank correlation coefficient (ρ) are introduced to evaluate linear and non-linear relationship(Fuchs, 2010), respectively. The results of factor correlation analysis for two predicting variables TN (5, t+1) are shown in Table 1.

Table 1. Input and output variables for TN model.

Model	Time Lag	Output	Input Variable
TN-1	1 week	TN(5)	TN(2,3,5,6,8), DO(4)
TN-2	1 week	TN(5)	PH(1,2,4,5,6,7,8), TN(2,3,4,5,6,7,8), C(1,5,8), DO(1,2,4), PPI(2,4), NH_3-N(1)
TN-3	1 week	TN(5)	All candidate variables

Note: Values in the brackets represent the number of GCL monitoring station.

RESULTS

All three models are able to capture the relationship between TN and input variables. In the training dataset, the correlation coefficient values between predictions and observations are 0.991 of TN-1 and 1.000 of both TN-2 and TN-3. Also, the structure of ANNs have been simplified and optimized, attributed to the introduction of factor correla-tion analysis. Moreover, the comparison of modeling results be-tween TN-1 and TN-2 also determines the selection range of correlation coefficient. The training and validation results of three TN prediction models are shown in Figure 2.

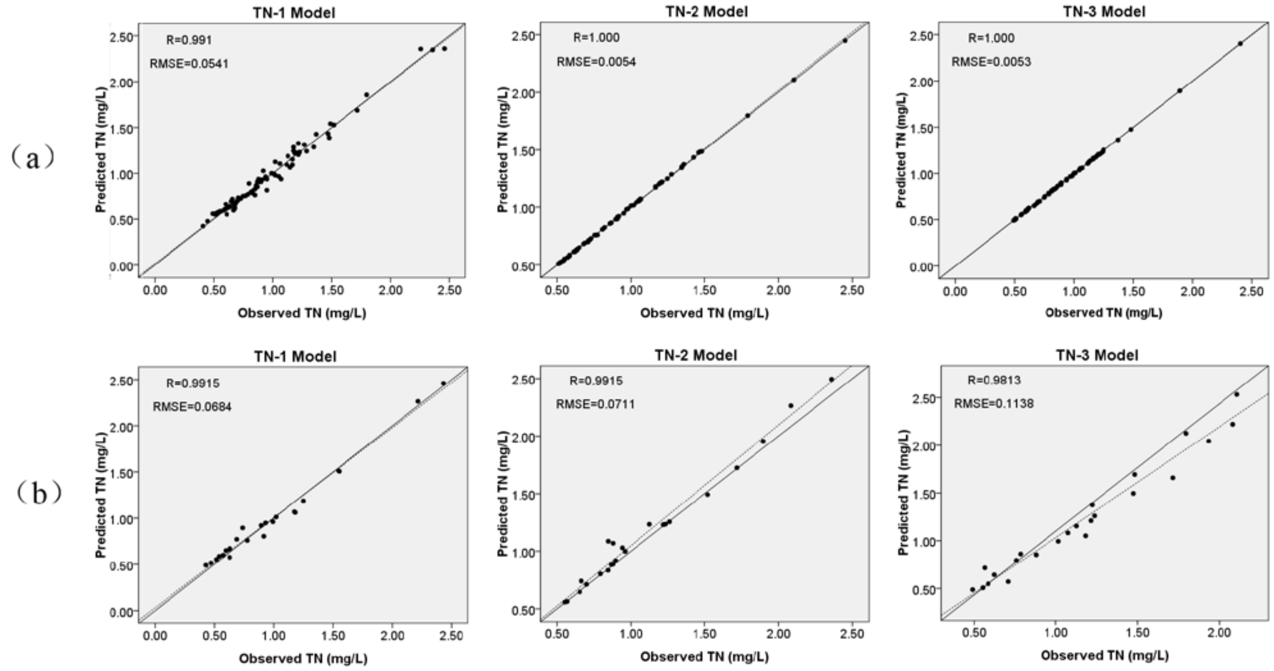

Figure 2. Training and validation results of three TN prediction models: (a) training and (b) validation sets.

DISSCUSSION AND CONCLUSIONS

Compare to similar studies, the modeling performance appears to be relatively better in this research. We consider that there are three main reasons. Initially, the investigation and analysis of the pollution influx and efflux approaches, which is often neglected or replaced by simple empirical judgement in similar research, plays a critical role in the development of sampling strategy (Kim, 2012). On the basis of that, the data collected could cover most governing factors affecting TN concentration at the center of Gucheng Lake, so the output and input variables are highly-correlated in this research. Secondly, the eutrophication aggravation and water quality deterioration process of Gucheng Lake is typical and representative during the sampling collection period. Therefore, the modeling data in this study contains vital information about eutrophication developing mechanisms of the lake. In addition, factor correlation analysis with selecting range of [-1.000, -0.500] and [0.500, 1.000], determining those variables which have great contribution to the target outputs, could optimize both modeling structure and prediction accuracy by identifying and eliminating surplus data among potential variables. Therefore, the models possess good generalization capability and are able to stimulate the relationship between inputs and outputs, and, finally, produce satisfactory results.

REFERENCES

Carpenter, S.R., Caraco, N.F., Correll, D L and et al. (1998) Nonpoint pollution of surface waters with phosphorus and nitrogen. Ecological Applications, 8, 559-568.

Fuchs, H.L., and Franks, P.J.S. (2010) Plankton community properties determined by nutrients and size-selective feeding. Marine Ecology Progress Series, 413, 1-15.

Kim, R., Loucks, D., and Stedinger, J. (2012) Artificial Neural Network Models of Watershed Nutrient Loading. Water Resources Management, 26, 2781-2797.

Melesse, A.M., Krishnaswamy, J., and Zhang, K.. (2008) Modeling Coastal Eutrophication at Florida Bay using Neural Networks. Journal of Coastal Research, 190-196.

Nixon, S.W. (1995) Coastal Marine Eutrophication - a Definition, Social Causes, and Future Concerns. Ophelia, 41, 199-219.