Artificial Neural Networks in the Prediction and Assessment for Water Quality: A Review

Yingyi Chen 1,2,3,*, Xiaomin Fang 1,2,3, Ling Yang 1,2,3, Yeqi Liu 1,2,3, Chuanyang Gong 1,2,3 and Yuqi Di 2,3,4

1College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China
2Key Laboratory of Agricultural Information Acquisition Technology, Ministry of Agriculture, Beijing, 100081, P.R. China
3Beijing Engineering and Technology Research Centre for Internet of Things in Agriculture, Beijing, 100097, P.R. China
4Information Technology Academy, Jilin Agricultural University, Changchun, Jilin, 130118, China

* Corresponding author’s e-mail: chenyingyi@cau.edu.cn

Abstract. Water is one of the main elements of the environment, which determines the existence of life on the earth, such as humans, aquatic animals, and plants. In order to control the water quality environment more efficiently and intelligently, numerous water quality models have been developed for predicting and evaluating water quality accurately and intelligently. In order to control the water quality environment more effectively and intelligently, artificial neural network (ANN) and the hybrid models that contain it are applied to accurately and intelligently predict and evaluate water quality, improving the reliability and assessment capabilities of water quality prediction. Therefore, this paper is a literature review aimed at analysing and comparing the characteristics and applications of existing artificial neural network models. According to the direction of information transmission in the network, we divide them into feed-forward networks and recurrent networks. In addition, we compare the pros and cons of each model. Our analysis provides guidance for model improvement in future research. Moreover, these models can be applied to aquaculture in the future to promote their development.

1. Introduction

Water is one of the main elements of the environment which determine the existence of life on the earth, such as humans, aquatic animals, and plants. However, the amount of wastewater from industrial, agricultural and domestic sewage is increasing sharply every year, and rivers are so polluted that water quality monitoring and management are imminent. At present, water quality problems mainly include maintaining oxygen levels, eutrophication (excessive growth of algae caused by excessive nitrogen and phosphorus in water), pH and heavy metal pollution. In response to these problems, the water quality models simulate the migration and transformation process of pollutants, identify the temporal and spatial distribution of pollutants and water quality parameters, and provide powerful technical and methodological support for water quality prediction, management and planning decisions. Therefore, establishing a water quality model is the most important task in solving water quality problems.
In order to control the water quality problems within the water environmental load capacity, a lot of models were built to evaluate or simulate the water quality in the last few years. The first water quality model is the S-P model proposed by Streeter and Phelps to control river pollution in Ohio, USA[1]. Since then, many models were put forward based on the different amounts of the data and different tools. From these two principles, we can divide the models into three phases. In the first stage, the models were one-dimensional steady-state models based on small data which meant that the amount of data was small. They mainly focused on oxygen balance. Later, the scholars studied the effects of pollutants on the water environment and found that the same pollutant had different effects on the water environment due to differences in physical state and chemical reactions. We call it the second stage. In this stage, to explore the relationship among the parameters, multi-media and multi-dimensional dynamic ecological models were proposed. The models focused on the simulation of the actual environment to estimate the value of a parameter in a certain situation. With the birth of machine learning theory, scholars only consider the key influence factors of target parameters by using intelligent algorithms to extract features from data. And based on the improvement of computer computing capability and the increase of data volume, researchers used artificial neural networks (ANN) to improve the reliability and evaluation capabilities. In this stage, these models have all been applied and played a major role in water environment management. The artificial neural network has exerted great advantages and solved the problems of poor adaptability and low efficiency of the traditional model. More importantly, the aquaculture industry is beginning to use modeling to analyze and control the water environment, which leads the intelligent development of aquaculture, and greatly increasing aquaculture production.

This paper focuses on the application of artificial neural network models and their hybrid model in water quality prediction and assessment. Starting from the network architecture, the artificial neural networks are divided into two types: feed-forward networks and recurrent networks. It proves the advantages of new technologies and their minimal application in aquaculture, and has great research space by comparing and evaluating the mean absolute error (MAE), mean squared error (MSE), root mean square error (RMSE) and squared correlation coefficient (R^2) and other criteria for each model. And it puts forward suggestions and prospects that these new techniques and methods can be applied to aquaculture water quality prediction in the following studies. Finally, the significance of these models and the paper is summarized. This paper provides a good guide to the future research on water quality prediction and assessment.

2. Classification

Artificial Neural Network (ANN) is a research hotspot in the field of artificial intelligence since the 1980s. It abstracts the human brain neuron network from the perspective of information processing, establishes a simple model, and forms different networks according to different connection methods. A neural network is an operational model consisting of a large number of neurons connected to each other. The network architecture resolves the integrated structure and information flow of the ANN models. Traditionally, the ANN architecture is divided into feed-forward and recurrent networks (figure 1).

In a feed-forward network, information is propagated only from the input layer to the output layer. The most common feed-forward model architectures are multilayer perceptron (MLP), generalized regression neural network (GRNN), radial basis function (RBF) networks, convolutional neural network (CNN) and so on. In a recurrent network, information can be propagated not only through the feed-back loop in the forward direction, but also through the output layer neurons to feed back the input and/or hidden layer neurons. The existence of a feed-back mechanism in the recurrent network makes it easier for neural networks to model highly dynamic systems with time delays. Back propagation neural network (BPNN) is the simplest cyclic network. In addition, it includes long- and short-term memory network (LSTM), extreme learning machine (ELM), and so on.

Due to the complex interaction between water quality parameters, most of the water quality parameters are nonlinear, temporal and spatial, and unstable. These add a lot of uncertainty to the
prediction and evaluation of water quality parameters. Therefore, relying solely on artificial neural network models may not capture the complex nature of environmental and hydrological systems. Therefore, many hybrid models have been proposed to improve prediction and evaluation accuracy. The original artificial neural network model is improved mainly from two aspects: data input and parameter optimization of the model itself.

![Figure 1. Taxonomy of model architectures](image)

### 3. Applications

#### 3.1. Feed-forward networks

The MLPNN model consists of three parts: the input layer, one or more hidden layers, and an output layer. The number of hidden layers in an MLPNN model depends on the complexity of the process being modeled. Each input is converted to all neurons in the first hidden layer. Each neuron in the first hidden layer converts its input signal into an output signal that is sent to the next layer in the ANN architecture. In this way, the original input vector travels forward through the ANN. The last hidden layer sends its output signal to the output layer. The final output layer decodes the signal into a final response to the original stimulus (input). Haykin[2] provided detailed information about the MLP architecture. The MLP architecture has been shown to approximate complex nonlinear functions to arbitrary precision[3]. Its original model was used in water quality parameter prediction.

RBFNN is a single hidden layer feedforward network based on function approximation theory[4]. The hidden layer adopts the radial basis function of the local response as the excitation function, having excellent characteristics such as simple structure, fast training speed and independent of initial weight. Hossam Adel Zaqoot et al[5] experimented with RBFNN to predict the dissolved oxygen concentration in the Mediterranean Sea water along Gaza within two weeks. Water temperature, wind speed, turbidity, pH, and conductivity are treated as inputs to the network. It turned out that RBFNN has good adaptability and wide applicability for the prediction of dissolved oxygen, but it is slightly inferior to the MLP improved by Levenberg-Marquardt optimization (LM)[6-7].

The GRNN was proposed by American scholar Specht in 1991[8]. It’s a four-layer forward neural network (input layer, mode layer, summation layer, and output layer) in the RBFNN. It has strong nonlinear mapping capability and flexible network structure, and has good fault tolerance and robustness. In view of the complexity of pond dissolved oxygen concentration affected by many factors, Shi Pei et al[9], chose to construct a prediction model for dissolved oxygen based on GRNN and BPNN (see section 3.2). Through experimental comparison, the absolute relative error of the two neural networks for the prediction of dissolved oxygen in water is 7.48% and 22.39%, respectively. It can be seen that both neural networks can be applied to the prediction of dissolved oxygen in aquaculture, but GRNN is more advantageous.

Extreme Learning Machine (ELM) is a kind of machine learning algorithm based on feedforward neuron network. Its main feature is that the hidden layer node parameters can be random or artificially given without adjustment. Xinfei Li et al[10], established a pond dissolved oxygen prediction model for the ELM neural network and improved dissolved oxygen distribution based on artificial push flow. The model was used to predict the dissolved oxygen concentration of different voltage pumps. The mean square error was 0.0394 and the correlation coefficient R2 was 0.9823, which were higher than BPNN.
In Xuxiang Ta’s published article[11], a simplified inverse understanding of the Convolutional Neural Network (CNN) prediction model is proposed to solve the dissolved oxygen prediction problem. The model multiplied the input vector containing the normalized input data by its transpose and simulated the input matrix of the original CNN. Experiments showed that backward understanding of CNN performance was better than BPNN.

3.2. Recurrent networks
As the simplest recurrent network, BPNN’s work process is divided into a study period and a work period, the study period is divided into two phases: (1) Forward transfer. The input information flow is processed from the input layer, through the hidden layer to the output layer, and the actual output value of each neuron point is calculated. (2) Error back propagation. Calculate the error between the actual output of the network and the expected value of the training sample. If the error does not reach the allowable value, determine the adjustment amount of the weight according to the error, and modify the connection weight of each layer of the neuron node from the back to the front. The working period begins after a large number of sample trainings during the learning period, during which only the positive propagation of the input information is available.

Traditional BPNN has been applied to water quality management many times[12-15]. For example, in order to verify the validity of the BPNN in river water quality prediction, Zhihong Zou and Xueliang Wang[16] used the water quality data of the upstream known river sections to predict the water quality of the downstream detection section of the river. The conclusion showed that BPNN is suitable for long-distance prediction of short river sections.

Long-term memory (LSTM) neural network, as an improved model of recurrent neural network (RNN), can effectively solve the problems of gradient disappearance and gradient explosion during RNN training and greatly improve the accuracy of RNN. Yingyi Chen et al[17]. proposed a model for predicting dissolved oxygen in aquaculture based on principal component analysis (PCA) and long short-term memory (LSTM). Compared with others, the LSTM had the highest prediction accuracy for dissolved oxygen.

In Shuangyin Liu’s study[18], the Elman neural network was used to predict the dissolved oxygen concentration in the Hyriopsis Cumingii pond. It is observed that the Elman NN method with 7 or 8 hidden layer nodes is a better architecture of the DO model, but the accuracy of the DO model is not good. And predictive test data has obvious interference in the front and back, which is an inherent problem of Elman neural network.

3.3. Hybrid models
Artificial neural networks have strong ability to deal with nonlinear relationships, but studies to date have shown that neural networks have some intractable defects, such as the difficulty in determining the network structure, local optimization, and easy to learn and generalize. Therefore, the application effect of the neural network is reduced to some extent. It is these shortcomings that have led scholars to further research, combining traditional neural networks with various optimization algorithms to make up for these shortcomings, making the neural network's advantages in prediction more prominent.

3.3.1. Data input
In response to the complexity of predicting Water quality parameter, Jianxun He et al[19]. proposed a partial mutual information (PMI) input selection algorithm (which identifies input variables in a stepwise manner) for selecting a set of input variables for the development of MLPNN. The method exhibited superior performance over methods including ANNs that were input by using partial correlation and all potential input fed in flow modeling.

Yingyi Chen et al[20], based on RBFNN, combined with subtractive clustering (SC) method and K-means clustering method to improve the accuracy of obtaining the number of hidden units in RBF neural network.
In order to further improve the accuracy and solve the problem of poor robustness of traditional methods, Xiaohong Peng et al.[21]. proposed a dissolved oxygen prediction model for aquaculture based on principal component analysis (PCA) and GRNN. The model established the principal components of the PCA-based aquatic eco-environmental indicators by collecting the principal components of the PCA-based aquatic eco-environmental indicators. The error of the prediction result is no more than 5%, which proved that the model was feasible.

Chen Li et al.[22]. raised a hybrid model based on multi-scale features, which used the ensemble empirical mode decomposition (EEMD)[23]. First, the original DO data set was broken down into several components by EEMD. Second, these components were used to reconstruct four terms, including high frequency, medium, low, and trend terms. Third, using appropriate algorithms for different components of the data according to their characteristics.

3.3.2. Parameter optimization

Firefly algorithm (FFA), as an optimization tool for artificial intelligence models, has recently entered the field of predictive modeling[24-25], which is mixed with MLPNN[26]. The hybrid model was nearly 13% more accurate for DO prediction than the traditional MLPNN model.

In fact, the conventional RBFNN and BPNN often lead to longer training time and fall into local minimum easily. In this respect, many scholars have improved the two models. To avoid the neural networks entering the local small and improve the prediction accuracy as the goal, Hu Xuemei[27] and Bing Li[28] employ genetic algorithm (GA) to train the parameters of RBFNN. Experiments show that the GA-RBF has higher prediction accuracy and can fully realize the nonlinear system modeling. There was also a self-organizing radial basis function neural network model predictive control (SORBF-MPC) method which was used to control the DO concentration of wastewater treatment plants[29]. The SORBF can dynamically change its structure to maintain prediction accuracy. Hidden nodes in the RBF neural network can be added or deleted online based on node activity and mutual information (MI) to achieve proper network complexity and necessary dynamics, which can greatly improve the accuracy of the model without adjusting the weight of the RBF. For BPNN, algorithm particle swarm optimization (PSO)[30] and harmony algorithm[31] were used to adjust the weights in the network.

For the problem of slow speed, Jinting Ding et al. proposed to use the fuzzy control theory to train the feedforward artificial neural network with adaptive variable step size algorithm to reduce the training time of BPNN and improve the convergence efficiency and network stability[32]. BPNN often used several modified training functions, such as trainlm to improve the model. These functions have adaptive learning rates. Trainlm can update weights and bias values based on LM. The trainlm function shows the fastest calculation speed because of the use of LM.

4. Conclusions and prospects

As can be seen from the above review, in the prediction and evaluation of water quality parameters, the feedback network is suitable for time series data, and the feedforward network is suitable for ordinary data, and the performance of the hybrid model is better than the original model. In all the applications listed in this paper, the accuracy of the model is basically improved from the following two aspects: (1) Input of data, including calculation of correlation to screen the influence factor and data preprocessing. (2) Optimization of the parameters of the model itself. Using various optimization algorithms, find the most suitable weight combination of the model and get the optimal model structure. Various improvement methods have also achieved remarkable results. For future research, we can start from the following aspects.

• (1) Screening of impact factors. Collect more parameters, use traditional neural networks to search their effects on parameters prediction results, and remove the unrelated influence factors to reduce the redundant data.

• (2) Preprocessing of data. In addition to the simplest linear interpolation and mean smoothing methods, data cleaning algorithms can also be used in the pretreatment of related data, such as
Bayesian networks, rough set theory, etc. Most of these methods need to judge the similarity between missing records and complete records, which is the core issue.

- **(3) Selection of model parameters.** The particle swarm optimization algorithm (PSO) mentioned above is in the category of group intelligence algorithms. Group intelligence algorithms are derived from simulations of natural biota. It is a heuristic algorithm that achieves optimization by artificial algorithm simulation of the process of searching for food or exchanging information. These algorithms are all directional methods based on probability search[33]. The group intelligence algorithms proposed in recent years, such as the ant colony optimization algorithm (ACA)[34], the Moth-Flame Optimization (MFO) algorithm [35], Ant Lion Optimizer (ALO)[36], Dragonfly Algorithm (DA)[37] and Whale Optimization Algorithm (WOA)[38], should be applied to the establishment of WQMs. And conduct a series of comparative experiments to find the best model structure.

- So far, most scholars have studied the modeling of water quality in rivers, lakes and sewage, and have neglected the aquaculture sector. The development of the aquatic industry urgently requires the joining of new technologies to become intelligent. At this stage, scholars are most concerned about how to improve existing models to improve the accuracy of water quality parameter simulation and prediction. With the birth of more and more optimization algorithms, the simulation and prediction accuracy of hybrid models is getting higher and higher. I believe that on the basis of predecessors, scholars can develop more perfect models and apply them to aquaculture in the future.

**Acknowledgments**

This research is supported by the Science & Technology Program of Beijing “Research and Demonstration of technologies equipment capable of intelligent control for large-scale healthy cultivation of freshwater fish” (No. Z171100001517016) & the Chinese and British Intergovernmental Science and Technology Cooperation Program “Advancing digital precision aquaculture in China (ADPAC)” (No. 2017YFE0122100) & the Shandong Province key Research & Development Program “Research and Demonstration of accurate monitoring and controlling technologies for environment of vegetable in facility” (No.2017CXGC0201).

**References**

[1] STREETER, H. W., PHELPS, E. B., (1925) A STUDY OF THE POLLUTION AND NATURAL PURIFICATION OF THE OHIO RIVER. Public Health Service: Washington DC, USA Public Health Bulletin 146.

[2] Haykin, S., (1999) Neural networks: A comprehensive foundation. Tom Robbins, Hamilton, Ontario, Canada.

[3] HORNIK, K., (1991) Approximation Capabilities of Multilayer Feedforward Networks. Neural Networks, 4: 251-257.

[4] Broomhead, D. S., Lowe, D., (1988) Multivariable Functional Interpolation and Adaptive Networks. Complex Systems, 2: 321-355.

[5] Zaqoot, H. A., Ansari, A. K., Unar, M. A., et al., (2009) Provided for non-commercial research and educational use only. Not for reproduction or distribution or commercial use. Water Science & Technology, 60: 3051-3059.

[6] Maier, H. R., Jain, A., Dandy, G. C., et al., (2010) Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environmental Modelling & Software, 25: 891-909.

[7] Palani, S., Liong, S., Tkalić, P., (2008) An ANN application for water quality forecasting. Marine Pollution Bulletin, 56: 1586-1597.

[8] Specht, D. F., (1991) A general regression neural network. IEEE Trans Neural Netw, 2: 568-76.

[9] Pei, S., Yongming, Y., Hongyan, Z., et al., (2017) Application of GRNN and Elman Neural Network in Prediction of Dissolved Oxygen in Water. Jiangsu Agricultural Sciences, 45: 217-221.
[10] Li, X., Ai, J., Lin, C., et al., (2018) Prediction model of dissolved oxygen in ponds based on ELM neural network. IOP Conference Series: Earth and Environmental Science, 121: 22003.

[11] Ta, X., Wei, Y., (2018) Research on a dissolved oxygen prediction method for recirculating aquaculture systems based on a convolution neural network. Computers and Electronics in Agriculture, 145: 302-310.

[12] Guo, Q., He, Z., Li, L., et al., (2011) Application of BP neural network model for prediction of water pollutants concentration in Taihu Lake. Journal of Southern Agriculture, 42: 1303-1306.

[13] Wu, G., Qin, J., Wan, Q., (2013) Water Quality Evaluation of Xiangjiang Estuary Based on BP Neural Network Model. Journal of Natural Science of Hunan Normal University, 36: 92-95.

[14] Huang, Y., Ji, D., (2008) Experimental study on seawater-pipeline internal corrosion monitoring system. Sensors and Actuators B: Chemical, 135: 375-380.

[15] Xiao, Z., Peng, L., Chen, Y., et al., (2017) The Dissolved Oxygen Prediction Method Based on Neural Network. Complexity, 2017: 1-6.

[16] Zou, Z., Wang, X., (2007) The errors analysis for river water quality prediction based-on BP-modeling. Acta Scientiae Circumstantia, 27: 1038-1042.

[17] Yingyi, C., Qianqian, C., Xiaomin, F., et al., (2018) Principal component analysis and long short-term memory neural network for predicting dissolved oxygen in water for aquaculture. Transactions of the Chinese Society of Agricultural Engineering, 34: 183-191.

[18] Liu, S., Yan, M., Tai, H., et al., (2011) Prediction of Dissolved Oxygen Content in Aquaculture of Hyriopsis Cumingii Using Elman Neural Network. In: Computer and Computing Technologies in Agriculture V. Beijing, China. pp.508-518.

[19] He, J., Valeo, C., Chu, A., et al., (2011) Prediction of event-based stormwater runoff quantity and quality by ANNs developed using PMI-based input selection. Journal of Hydrology, 400: 10-23.

[20] Chen, Y., Yu, H., Cheng, Y., et al., (2018) A hybrid intelligent method for three-dimensional short-term prediction of dissolved oxygen content in aquaculture. PLOS ONE, 13: e0192456.

[21] Peng, X., Chen, L., Yu, Y., et al., (2016) PCA-GRNN-GA Based PH Value Prediction Model Applied in Penaeus Orientalis Culture. In: 2016 6th International Conference on Digital Home (ICDH) (2016). Guangzhou, Guangdong, China. pp.227-232.

[22] Li, C., Li, Z., Wu, J., et al., (2018) A hybrid model for dissolved oxygen prediction in aquaculture based on multi-scale features. Information Processing in Agriculture, 5: 11-20.

[23] Wu, Z., Huang, N. E., (2009) Ensemble Empirical Mode Decomposition: : A Noise-assisted Data Analysis Method. Advances in Adaptive Data Analysis, 1: 1-41.

[24] Ghorbani, M. A., Shamshirband, S., Zare Haghi, D., et al., (2017) Application of firefly algorithm-based support vector machines for prediction of field capacity and permanent wilting point. Soil and Tillage Research, 172: 32-38.

[25] Yang, X., (2010) Firefly Algorithm, Stochastic Test Functions and Design Optimisation. International Journal of Bio-Inspired Computation, 2: 78-84.

[26] Raheli, B., Aalami, M. T., El-Shafie, A., et al., (2017) Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River. Environmental Earth Sciences, 76:

[27] Xuemei, H., Yingzhan, H., Xingzhi, Y., (2011) The Soft Measure Model of Dissolved Oxygen Based on RBF Network in Ponds. In: 2011 Fourth International Conference on Information and Computing. Phuket Island, Thailand. pp.38 - 41.

[28] Li, B., Cong, L., Zhang, W., (2013) Research on Optimized RBF Neural Network based on GA for Sewage Treatment. In: 2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics. Hangzhou, China.

[29] Han, H., Qiao, J., Chen, Q., (2012) Model predictive control of dissolved oxygen concentration based on a self-organizing RBF neural network. Control Engineering Practice, 20: 465-476.
[30] Li, M., Wu, W., Chen, B., et al., (2017) Water Quality Evaluation Using Back Propagation Artificial Neural Network Based on Self-Adaptive Particle Swarm Optimization Algorithm and Chaos Theory. Computational Water, Energy, and Environmental Engineering, 6: 229-242.

[31] Yingbo, W., Lin, W., Deng, D., et al., (2012) Harmony Search Algorithm Optimized BP Network and Application in Water's Quality Evaluation. Computer Measurement & Control, 20: 1931-1933.

[32] Jinting, D., Zelin, Z., Min, H., (2017) Penaeus vannamei aquaculture water quality prediction based on the improved back propagation neural network. Journal of Zhejiang University (Agric.&Life Sci.), 43: 128-136.

[33] Donglai, C., Weichen, Z., Chuncheng, L., (2015) Research Status and Development Trend of Group Intelligence Algorithm. INFORMATION & COMMUNICATIONS, 11: 38-39.

[34] M., D., (1992) Optimization, Learning and Natural Algorithms. Doctor, Politecnico di Milano, Milano, Italy.

[35] Mirjalili, S., (2015) Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. Knowledge-Based Systems, 89: 228-249.

[36] Mirjalili, S., (2015) The Ant Lion Optimizer. Advances in Engineering Software, 83: 80-98.

[37] Mirjalili, S., (2016) Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. Neural Computing and Applications, 27: 1053-1073.

[38] Mirjalili, S., Lewis, A., (2016) The Whale Optimization Algorithm. Advances in Engineering Software, 95: 51-67.