*Green Energy, Environment and Sustainable Development*
*C. Wang et al. (Eds.)*

*doi:10.3233/ATDE230306*

# Advances in Machine Learning for Water Quality Prediction and Prospects in Erhai Lake

Xinyu ZHOU[a,b] and Jing ZHANG[a,b,1]
[a] *Beijing Laboratory of Water Resources Security, Capital Normal University, Beijing 100048, China*
[b] *The Key Lab of Resource Environment and GIS of Beijing, Capital Normal University, Beijing 100048, China*

**Abstract.** Water quality research plays a pivotal role in addressing and mitigating water pollution issues, while accurate water quality forecasting is vital for safeguarding the ecological integrity of watersheds. In recent years, the integration of machine learning models into water quality prediction has garnered significant attention from scholars due to their inherent advantages, including adaptability, self-learning capabilities, high efficiency, and fault tolerance. This paper aims to provide a comprehensive overview of the application of machine learning techniques in water quality prediction models. Additionally, it critically examines the existing challenges and limitations associated with these prediction models. Furthermore, this study presents a forward-looking perspective on the potential utilization of machine learning approaches in forecasting water quality models specific to the Erhai Lake region. By consolidating and analyzing the available knowledge, this research endeavors to contribute to the advancement of machine learning-based water quality prediction methodologies in order to enhance the effectiveness and accuracy of future predictions in the Erhai Lake area.

**Keywords.** Machine learning, water quality prediction, remote sensing

## 1. Introduction

Water quality parameters serve as crucial indicators for assessing the ecological environment of water, reflecting its quality level and trends [1]. Research on water quality is essential for addressing and preventing water pollution. Protecting water quality is vital for ensuring human health, societal development, and ecological safety in inland waters [2]. Water quality research provides a scientific foundation for the formulation of water pollution prevention and control policies and the regulation of urban production and domestic sewage discharge.

In recent years, artificial intelligence technology, particularly machine learning methods, has been increasingly integrated into water quality remote sensing monitoring and prediction due to their ability to unravel complex natural relationships [3]. Remote sensing data has become the predominant data source in water quality models for river

[1] Corresponding Author, Jing ZHANG, Beijing Laboratory of Water Resources Security; The Key Lab of Resource Environment and GIS of Beijing, Capital Normal University, Beijing 100048, China; Email: 5607@cnu.edu.cn.

and lake basins and coastal areas. Compared to the complexity, data variety, and uncertainty of ecological mechanism models, remote sensing technology simplifies data-driven machine learning methods, making them more efficient and versatile.

The rapid development of the tourism industry around Erhai Lake in recent years has brought economic prosperity but also increased pollution to its water quality [4-6]. Consequently, numerous researchers have taken an interest in studying the water quality of Erhai Lake. Several scholars have recommended water quality prediction models, which have shown significant improvements in the water quality of Erhai Lake. This paper aims to explore machine learning methods for water quality prediction models and summarize the progress made in this field. Additionally, it discusses the challenges encountered during the water quality prediction process and presents the authors' perspectives on machine learning water quality prediction models. The study focuses on summarizing and classifying research methods employed by various scholars in the hope of obtaining highly accurate approaches to water quality prediction models.

## 2. Research Methodology

### *2.1. Principles of Water Quality Prediction*

Water quality prediction involves building a model based on historical measured data and using the model's input parameters to forecast changes in water quality. Two commonly used prediction models are Principle-Driven Models (PDM) and Data-Driven Models (DDM).

PDMs are typically based on expert knowledge, incorporating principles of mass and energy conservation, biological dynamics, hydrodynamics, water quality component interactions, and biochemical effects derived from complex ecological mechanism models such as EFDC, MIKE [7], SALMO [8], and SWAT [9]. However, PDMs are challenging to construct due to their lengthy and complex referencing processes, difficulty in unifying data types, and limited universality caused by their fixed structures.

In contrast to Principle-Driven Models (PDMs), Dissimilarity-based Data Mining (DDM) methods rely on the utilization of a substantial amount of water quality data within a learning model. These methods employ algorithms to iteratively adjust the model's parameters, facilitating the establishment of a mapping relationship between the input data and the corresponding predicted data. By harnessing the power of machine learning algorithms, DDM methods effectively capture the intricate non-linear associations that exist among various ecological indicators present in lakes and reservoirs. This is achieved by comparing actual data with model calculations, thereby extracting valuable information and revealing the underlying correlations among the diverse set of indicators. As a result, these DDM techniques yield significant advantages, including the reduction of computational time, cost, and errors associated with crucial tasks such as water quality categorization, prediction of water quality parameters, and estimation of water quality indices [10].

### *2.2. Machine Learning Methods*

According to a number of studies focusing on the water quality of Erhai Lake, researchers commonly employ machine learning methods such as Decision Tree (DT), Support Vector Machines (SVM), Neural Networks (NN), and Random Forests (RF) for water

quality prediction. These methods have been recognized for their effectiveness in achieving accurate predictions. However, each method possesses distinct advantages and disadvantages, which are summarized in Table 1. This comprehensive evaluation allows researchers and practitioners to make informed decisions regarding the selection and application of appropriate machine learning methods for water quality prediction in the context of Erhai Lake.

**Table 1.** Advantages and disadvantages of the four machine learning methods.

| Machine learning methods | Advantages | Disadvantages |
|---|---|---|
| DT | Simplicity and efficiency in handling big data, flexibility in dealing with various attribute types. | Proneness to overfitting, difficulty in predicting continuous fields with large sample data. |
| SVM | Better generalization performance, less prone to overfitting, good performance with limited data. | Slowness with large-scale training samples, sensitivity to missing data and parameters. |
| NN | High classification accuracy, powerful parallel processing and learning capabilities. | Requirement of a large number of parameters, excessive learning time without guaranteed goal achievement. |
| RF | Good model generalization and handling of high-dimensional data without feature selection. | Potential overfitting in noisy classification or regression problems. |

### *2.3. Current Status of Research*

Remote sensing data is widely used in water quality prediction models. For example, Tian [11] employed Sentinel-2 imagery to compare four machine learning algorithms, with XGBoost demonstrating superior performance in retrieving chlorophyll-a (Chl-a), dissolved oxygen (DO), and NH3-N from inland reservoirs. Xiao et al. [12] constructed a groundwater quality prediction model using BP neural network, RF, and SVM, highlighting the significant influence of different machine learning methods and lag period selection on predictability. Zhang [13] proposed a multinomial least squares vector regression model for water quality prediction, utilizing a committee of non-linear functions to map samples into a high-dimensional feature space and employing linear regression for fitting. This approach aims to capture complex relationships between water quality variables and predictors, enhancing prediction accuracy. Shi et al. [14] developed a wavelet analysis-based long and short memory neural network model (WA-LSTM) for water quality prediction, demonstrating improved accuracy and capturing data characteristics through wavelet analysis and LSTM. Nunes [15] explored machine learning models for forecasting Chl-a in reservoirs and examined its correlation with hydrological and meteorological variables using satellite data. The study revealed that conventional empirical correlation between Chl-a and phosphorus is not applicable to tropical reservoirs. Instead, mean surface temperature, water level, and surface solar radiation showed direct relationships with Chl-a, while water volume and mixed layer depth exhibited inverse relationships.

Machine learning techniques are extensively used in estimating water quality indices as well. Xia [16] determined SVM as the better prediction model for the Entropy Water Quality Index (EWQI) based on correlation analysis. Four different algorithms were then employed to optimize the SVM model, with Differential Evaluation and Grey Wolf Optimizer (DE-GWO) yielding significantly higher accuracy. Ahmed [17] explored various supervised machine learning algorithms for estimating the Water Quality Index

(WQI), finding that gradient boosting achieved the best results.

## 3. Study Area

### *3.1. Introduction to the Erhai Region*

Erhai Lake, the second-largest highland freshwater lake in Yunnan Province, boasts remarkable natural conditions and abundant water resources that have nurtured the Bai people and their distinctive culture in Dali. The Erhai Lake basin is situated between 99°32'E-100°27'E and 25°25'N-26°16'N, encompassing a basin area of 2565 km2 and a lake area of 251 km2, within the Lancang-Mekong River system [18]. Erhai Lake serves as a vital source of drinking water for Dali, providing water for daily activities, irrigation, climate regulation, biodiversity preservation, and eco-tourism. It is a cornerstone for the sustainable economic and social development of the entire watershed and Dali Prefecture, earning it the title of "Mother Lake" among the local people [19]. Figure 1 depicts the digital elevation distribution of Erhai Lake.

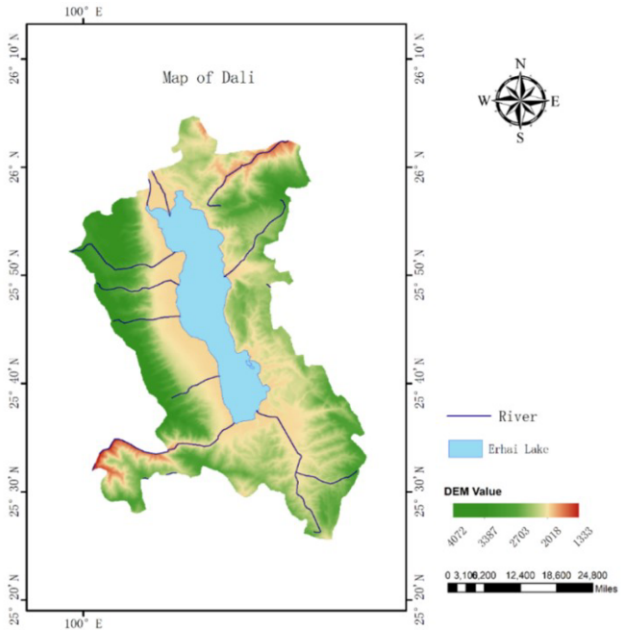


**Figure 1.** Study area.

### *3.2. Application of Machine Learning in Erhai Lake*

Yang [20] employed the NWI water body index method to extract Yilong Lake and Erhai Lake, constructing an artificial neural network-based model to accurately estimate historical lake surface water temperature with minimal errors. Zhang [21] proposed a comprehensive water quality prediction model using the RBF+NEW approach. Simulation results confirmed that the RBF+NEW model, compared to the radial basis network model, exhibited improved prediction accuracy and avoided the instability of

the BP neural network model.

### *3.3. Research Focuses on the Water Quality Prediction Model for Erhai Lake*

For the collected station data, vacant and abnormal values underwent data pre-processing. Analysis revealed significant fluctuations in Erhai Lake's water quality throughout the year, necessitating further investigation into its causal mechanisms. To provide early warnings of potential water quality deterioration, analyze water quality drivers, and facilitate timely measures, this study integrates the development of a water quality prediction model using DT, SVM, NN, and RF approaches. Furthermore, the application of a water quality prediction model for evaluation and optimization is explored. Figure 2 presents schematic diagrams of water quality indicators collected by the automatic monitoring station in Erhai Lake.

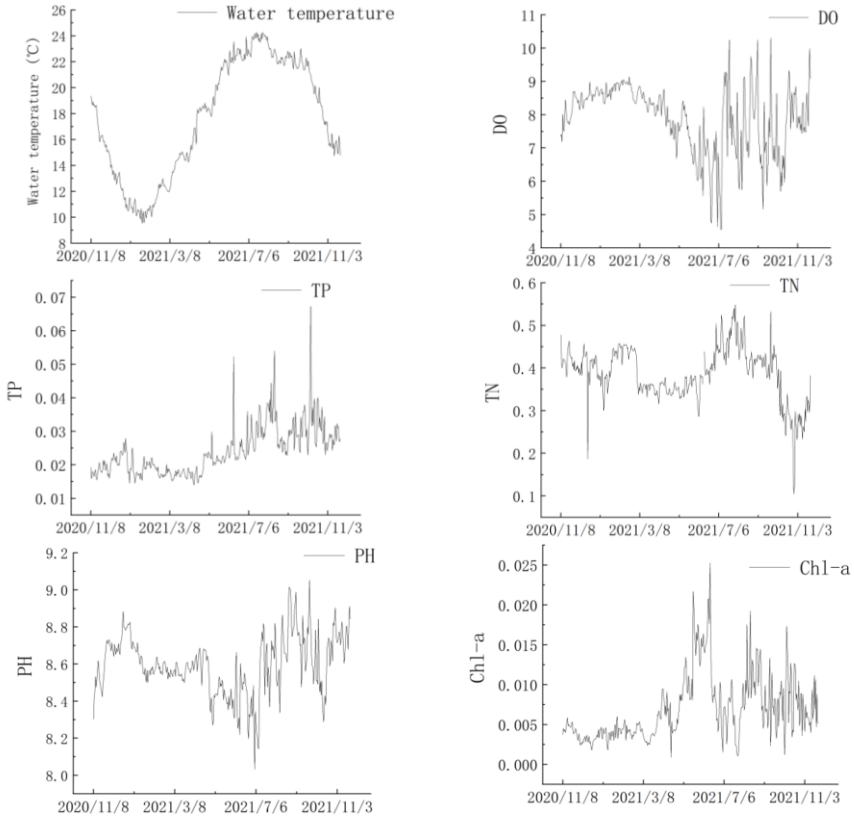


**Figure 2.** Schematic diagrams of water quality indicators.

## 4. Summary

Through the comparison of the four machine learning methods, it can be observed that these methods have achieved favorable results in water quality prediction within the Erhai study area, contributing to the development of water quality prediction models.

However, some challenges persist: (1) The data source for water quality prediction models is remote sensing data, which requires improvements in spatial resolution, radiation resolution, and uncertainties in data transmission and interaction processes. These non-human factors can result in missing and abnormal data. (2) Water quality characteristics are influenced by various factors, necessitating further exploration of how to include additional variables to enhance model prediction accuracy.

## References


[1] Wang SM, Qin BQ. Research progress on remote sensing monitoring of lake water quality parameters. Huanjing Kexue. 2023;44(3):1228-1243.

[2] Li X, Li Y, Li GJ. A scientometric review of the research on the impacts of climate change on water quality during 1998-2018. Environmental Science and Pollution Research. 2020;27(13):14322-14341.

[3] Peterson KT, Sagan V, Sloan JJ. Deep learning-based water quality estimation and anomaly detection using Landsat-8/Sentinel-2 virtual constellation and cloud computing. Giscience & Remote Sensing. 2020;57(4):510-525.

[4] Lin SS, et al. Assessment andmanagement of lake eutrophication: A case study in Lake Erhai, China. Science of the Total Environment. 2021;751.

[5] Yan CZ, Lu X, Zhao XL. Protection and sustainable utilisation of water resources in Lake Erhai basin. International Journal of Sustainable Development and World Ecology. 2008;15(4):357-361.

[6] Zhang Z, et al. Has government water protection policy taken effect on preventing harmful algal blooms in erhai lake?. IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Yokohama, Japan. 2019.

[7] Zhang XQ, et al. Simulation of diffuse source polluted water environment based on MIKE21: A case study of the urban section of the Bai river. Water Practice and Technology. 2022;17(9):1893-1913.

[8] Cetin L, Zhang B, Recknagel F. Process-based Simulation Library SALMO-OO for Lake Ecosystems. International Congress on Modelling and Simulation (MODSIM05). Melbourne, Australia. 2005.

[9] Anagnostou E, Gianni A, Zacharias I. Ecological modeling and eutrophication: A review. Natural Resource Modeling. 2017;30(3).

[10] Malek NHA, et al. Prediction of water quality classification of the Kelantan River Basin, Malaysia, using machine learning techniques. Water. 2022;14(7).

[11] Tian S, et al. Remote sensing retrieval of inland water quality parameters using Sentinel-2 and multiple machine learning algorithms. Environmental Science and Pollution Research. 2023;30(7):18617-18630.

[12] Jayaraman P, Nagarajan KK, Partheeban P. A review on artificial intelligence algorithms and machine learning to predict the quality of groundwater for irrigation purposes. 2022 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI). 2022;1-8.

[13] Zhang XZ, Yuan CG. Predict water quality based on multiple kernel least squares support vector regression and genetic algorithm. 12th International Conference on Control, Automation, Robotics and Vision (ICARCV). Guangzhou, China. 2012.

[14] Qing-yi SHI, et al. Evaluation and prediction of water quality of hongze lake based on machine learning method. China Rural Water and Hydropower. 2021;0(12)53-59.

[15] Carvalho TMN, Neto IEL, Souza FD. Uncovering the influence of hydrological and climate variables in chlorophyll-A concentration in tropical reservoirs with machine learning. Environmental Science and Pollution Research. 2022;29(49):74967-74982.

[16] Xia JJ, Zeng J. Environmental factors assisted the evaluation of entropy water quality indices with efficient machine learning technique. Water Resources Management. 2022;36(6):2045-2060.

[17] Ahmed U, et al. Efficient water quality prediction using supervised machine learning. Water. 2019;11(11).

[18] Yang J, et al. Monitoring of organochlorine pesticides using PFU csystems in Yunnan lakes and rivers, China. Chemosphere. 2007;66(2):219-225.

[19] Yang GH, Li ZX, Fan CQ. The effect of ecological rehabilitation of the Erhai lakeside on Odonata species richness and abundance. Aquatic Insects. 2017;38(4):231-238.

[20] Yang JY. Estimation and Change Analysis of Surface Water Temperature of Nine Lakes in Yunnan Based on Machine Learning Algorithm. Yunnan Normal University. 2021.

[21] Zhang Y. Research on Water Quality Prediction Algorithm Based on the Erhai Sea. Kunming University of Technology. 2018.