Statistics and Application 统计学与应用, 2026, 15(2), 179-187
Published Online February 2026 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2026.152045

Hans汉斯

基于机器学习模型的水质预测评估

王　杰

浙江师范大学地理与环境科学学院，浙江　金华

收稿日期：2026年1月25日；录用日期：2026年2月16日；发布日期：2026年2月26日

摘　要

随着城市化、工业化推进及人类活动强度增加，河流水环境问题日益突出，水质污染呈现多源性、时空异质性和复杂驱动性，水质预测成为水环境管理的重要需求。本研究为分析与预测主要地表水水质指标的时空分布特征，构建了随机森林(RF)、极端梯度提升(XGBoost)、决策树(DT)和自适应增强(AdaBoost)四种机器学习模型，并以决定系数(R^2)、均方根误差(RMSE)和平均绝对误差(MAE)作为评估指标。结果显示，四种模型的预测性能排序为RF > XGBoost > DT > AdaBoost，其中RF模型表现最优，其R^2、RMSE、MAE分别为0.985、0.042、0.024，具备极强的预测精度、稳定性及泛化能力，能有效捕捉水质指标间的复杂非线性关系。研究表明集成学习模型在处理非线性关系和抑制过拟合方面优势显著，RF模型可作为水质空间预测分析的最优模型，为精准开展水质动态模拟与水环境管理提供支撑。

关键词

河流水质，机器学习，水质预测，性能评估

Evaluation of Water Quality Prediction Based on Machine Learning Models

Jie Wang

College of Geography and Environmental Sciences, Zhejiang Normal University, Jinhua Zhejiang

Received: January 25, 2026; accepted: February 16, 2026; published: February 26, 2026

Abstract

With the advancement of urbanization and industrialization and the intensification of human activities, river water environments have been increasingly threatened. Water quality pollution is characterized by multiple sources, pronounced spatiotemporal heterogeneity, and complex driving mechanisms, making water quality prediction a critical requirement for effective water environment management. To analyze and predict the spatiotemporal distribution patterns of key surface water quality

文章引用: 王杰. 基于机器学习模型的水质预测评估[J]. 统计学与应用, 2026, 15(2): 179-187.
DOI: 10.12677/sa.2026.152045

indicators, this study developed four machine learning models, including Random Forest (RF), Extreme Gradient Boosting (XGBoost), Decision Tree (DT), and Adaptive Boosting (AdaBoost). The coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) were employed to evaluate model performance. The results indicate that the predictive performance of the four models followed the order: RF > XGBoost > DT > AdaBoost. Among them, the RF model exhibited the best performance, with R^2, RMSE, and MAE values of 0.985, 0.042, and 0.024, respectively, demonstrating superior prediction accuracy, stability, and generalization ability. The RF model effectively captured the complex nonlinear relationships among water quality indicators. The findings highlight the advantages of ensemble learning models in handling nonlinear processes and mitigating overfitting, suggesting that the RF model represents an optimal approach for spatial prediction and assessment of river water quality, thereby providing robust support for dynamic water quality simulation and water environment management.

Keywords

River Water Quality, Machine Learning, Water Quality Prediction, Performance Evaluation

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

河流水质是保障人类健康、生态系统平衡和经济发展的基础资源之一[1]。河流不仅承担供水、发电和蓄水等重要的社会经济功能，还在物质与水分循环、区域气候调节及生态系统稳定维系方面发挥着关键的生态环境作用，是支撑人类可持续发展和生态系统服务功能的重要基础[2]。然而，随着经济快速发展和人类活动强度的不断增加，过度开发与利用行为对河流水体造成的负面影响日益显现[3]，河流水质退化问题已在全球范围内普遍存在，成为当前水环境管理面临的重大挑战之一[4]。与此同时，土地利用强度的持续增加、快速城市化进程的推进、极端气候事件的频发以及温室效应加剧等因素，进一步加大了水资源调控与水环境管理的复杂性和不确定性[5]。

从全球尺度来看，尽管地球表面大部分被水体覆盖，水资源总量约为 14 亿 km^3，但其中可供人类直接利用的淡水资源极为有限，仅占地球水资源总量的约 0.3% [6] [7]。此外，全球水资源在空间分布上高度不均衡，不同国家和地区之间差异显著。水资源短缺已成为 21 世纪人类社会面临的第二大全球性问题[8] [9]。

实时监测系统的发展使高密度水质数据的获取成为可能[10]。然而，由于水质数据具有显著的非线性、动态性和非平稳性特征，河流水质的精确预测仍面临较大挑战[11]-[13]。这一问题还因监测站点之间复杂的时空依赖关系而进一步加剧[14]，而这种依赖性由水文连通性、空间拓扑结构以及外部扰动的共同作用所驱动，其中包括气象条件变化[15]和突发性污染事件[16]。鉴于水质演变过程中的高度不确定性以及对高精度时空预测需求的不断增长，亟需发展能够有效刻画水质数据动态性、非线性和非平稳特征的方法。

我国在水质预测模型研究方面起步相对较晚，但在引进和借鉴国外成熟模型的基础上，结合本国水环境特征对模型进行了改进与本地化应用，并取得了一定研究成果[17]。2010 年，李道亮等[18]将传统时间序列方法(自回归 AR 模型)与 BP 神经网络相结合，对海参养殖池溶解氧(DO)浓度进行预测，提升了预测精度。2018 年，周剑等[19]提出基于改进灰色关联分析(IGRA)与长短期记忆网络(LSTM)的水质预测方法，研究表明该方法能够充分挖掘水质指标之间的多元相关性和时间序列特征，其预测效果优于单一特

征或非序列模型。2022 年，杨林超等[20]构建了融合小波分解、自回归综合移动平均模型(ARIMA)和门控循环单元(GRU)的组合预测模型，在多项常规水质指标预测中表现出更高的精度；同时，采用集成学习模型 LightGBM 对水质评价等级进行预测，也展现出优良的模型性能。

近年来，随着人工智能技术的快速发展，数据驱动模型逐渐成为复杂非线性水环境系统建模的重要手段。随机森林(RF)、遗传规划(GP)、支持向量回归(SVM)、人工神经网络(ANN)、多元线性回归(MLR)以及自适应神经模糊推理系统(ANFIS)等方法，被广泛应用于河流水质模拟与预测研究中[21]。与传统统计模型或基于机理过程的模型相比，机器学习模型通常对专业机理知识依赖较少，模型开发和计算效率较高[22]，能够直接从数据中学习输入特征与输出目标之间的映射关系。此外，相较于线性回归模型，机器学习模型具有更强的非线性拟合能力，更容易捕捉水质指标之间复杂的动态关系，从而提高预测精度[23]。

2. 材料与方法

2.1. 研究材料

本研究所用水质数据来源于中国生态环境监测数据平台，涵盖研究区内多个固定监测断面。监测指标包括 DO、BOD_5、COD、COD_{Mn}、NH_3-N 和 TP。数据采样频率为月尺度，时间覆盖 2015~2022 年(表1)。在数据质量控制与清洗后，共获得 96 条有效样本，构成用于模型分析的完整数据集。在模型构建过程中，考虑到水质数据具有显著的时间序列特征，为避免信息泄露，采用基于时间顺序的样本划分方式，而非随机抽样。具体而言，按照时间先后顺序将前 80%的样本作为训练集，用于模型训练与参数优化，剩余后 20%的样本作为独立测试集，用于模型性能评估。该划分方式确保测试集数据在时间上晚于训练集，从而避免未来信息被用于模型训练，保证评估结果的客观性和可靠性。

Table 1. Descriptive statistics of concentrations of different water quality parameters
表 1. 不同水质指标浓度的描述性统计

水质指标	最大值(mg/L)	最小值(mg/L)	平均值(mg/L)
DO	20.7	0.87	7.12
BOD_5	11.8	0.2	1.75
COD	54.4	2	8.6
COD_{Mn}	9.32	0.3	2.42
NH_3-N	7.35	0.01	0.27
TP	0.546	0.001	0.07

2.2. 研究方法

水质指数(Water Quality Index, WQI)是一种用于综合评价水体整体水质状况的定量方法。该方法通过对多项水质参数进行标准化、加权和归一化处理，将复杂的监测数据转化为单一的综合指数，以反映水体环境质量的总体水平[24]。WQI 根据各水质指标对整体水质的重要性赋予不同权重，并通过加权标准化得分的累加计算得到最终的综合指数[25]。该方法在保证科学严谨性的同时，有效简化了多参数水质评价体系的复杂性，使不同时间、空间位置及监测断面之间的水质状况具有可比性[26]。WQI 的计算公式如下：

$$\mathrm{WQI} = \frac{\sum_{i=1}^{n} w_i q_i}{\sum_{i=1}^{n} w_i}$$

其中，WQI 表示综合水质指数，n 为参与评价的水质参数数量，w_i 为第 i 个水质参数的权重，反映其对整体水质状况的相对重要性，q_i 为第 i 个参数的水质分指数，用于表征实测值与相应水质标准之间的符合程度。各单项水质参数的分指数通常基于实测浓度与标准限值之间的关系，采用线性归一化公式计算：

$$q_i = \frac{C_i}{S_i} \times 100$$

其中，C_i 为第 i 个水质参数的实测浓度，S_i 为其对应的水质标准限值，依据《地表水环境质量标准》(GB 3838-2002)确定。当 $C_i \le S_i$ 时，q_i 值较低，表明水质状况良好；当 $C_i > S_i$ 时，q_i 值随之增大，反映水体污染程度的加重[25]。为减少极端值对综合评价结果的影响，通常对 q_i 设定上限值(如 100)，以提高 WQI 计算结果的稳定性和可比性[26]。

为分析与预测浙江省主要地表水水质指标的时空分布特征，本研究分别构建了独立的随机森林(RF)模型、极限梯度提升(XGBoost)、决策树(DT)和自适应增强(AdaBoost)。

RF 是由 Breiman 2001 [27]提出的一种监督机器学习算法，是一种典型的集成学习方法，通过融合多个决策树的预测结果来提升模型整体性能。其核心机制包括自助采样(Bootstrap Aggregating)，用于降低模型方差，以及特征子空间的随机选择，以提升各决策树之间的多样性[28]。该算法在处理非线性关系、高维数据及异常值时表现出较强的鲁棒性，同时具备评估输入特征重要性的能力[29] [30]。Boost 算法属于集成学习元算法(ensemble meta-algorithms)范畴，其核心目标在于提升弱学习器的预测能力[31]。在监督学习问题中，Boost 方法通过有效降低模型的偏差与方差，从而显著改善整体预测性能[32]。决策树(Decision Tree, DT)及其变体被归类为非参数监督学习方法[33]，该类方法通过将输入特征空间划分为若干互不重叠的区域，并为各区域分配独立的模型参数，从而实现对复杂非线性关系的刻画[34]。因此，决策树模型被广泛应用于分类、回归分析以及决策规则的可视化表达[35]。

为全面评估所构建模型的预测性能，本研究基于交叉验证过程中获得的折外(out-of-fold)预测结果进行了系统性评价。具体而言，采用了三种常用的回归评估指标：决定系数(R^2)、均方根误差(RMSE)和平均绝对误差(MAE) [36] [37]。其中，R^2 用于衡量模型对观测数据方差的解释程度，其取值范围为 0 至 1，值越大表示模型拟合效果越好；RMSE 与 MAE 反映预测值与观测值之间的偏差，其中 RMSE 对异常值更为敏感，而 MAE 则能够提供更稳健的平均误差估计[38]。在模型评估过程中，以逐年滚动的方式将每一年数据单独作为验证集，其余年份用于训练，从而更真实地反映模型在实际条件下对未来数据的预测能力。该评估策略有效避免了时间序列数据的泄漏问题，从而提升了模型性能指标在时序预测任务中的可解释性与可靠性。

3. 结果和讨论

3.1. 机器学习模型的性能评估

系统性评估结果表明，各模型预测性能的差异反映了其在复杂环境系统建模中的适应性差异。本研究中模型总体性能的排序为 RF > XGBoost> DT > AdaBoost (图 1 和表 2)。RF 模型在四项指标上均优于其他模型，这主要归因于其集成结构与随机特征选择机制。该机制能够有效捕捉多变量之间的非线性关系，显著提升模型对过拟合的抗性，并具备出色的泛化能力[29]。这些优势使得 RF 模型在高维度、多源及异质性的地表水系统建模中表现出极高的适用性。

利用 4 种机器学习模型对水质指数进行回归预测的结果如图 2 所示。随机森林模型的预测值和真实值拟合效果较好，在测试集的数据中，绝大部分预测值位于 y = x 直线的两侧。根据 4 种机器学习模型测试集预测值和真实值对比图(图 3)可以看出随机森林模型在基于水质参数的水质指数预测中具有良好的性能。

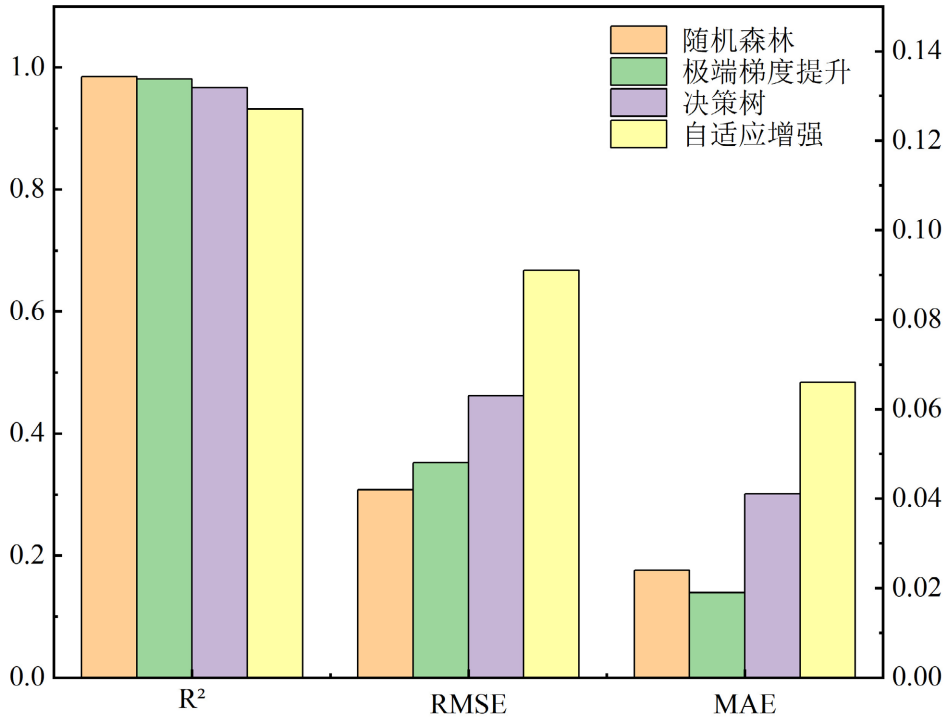

Figure 1. Comparison of prediction performance of four machine learning models on the test set
图 1. 4 种机器学习模型测试集预测性能对比

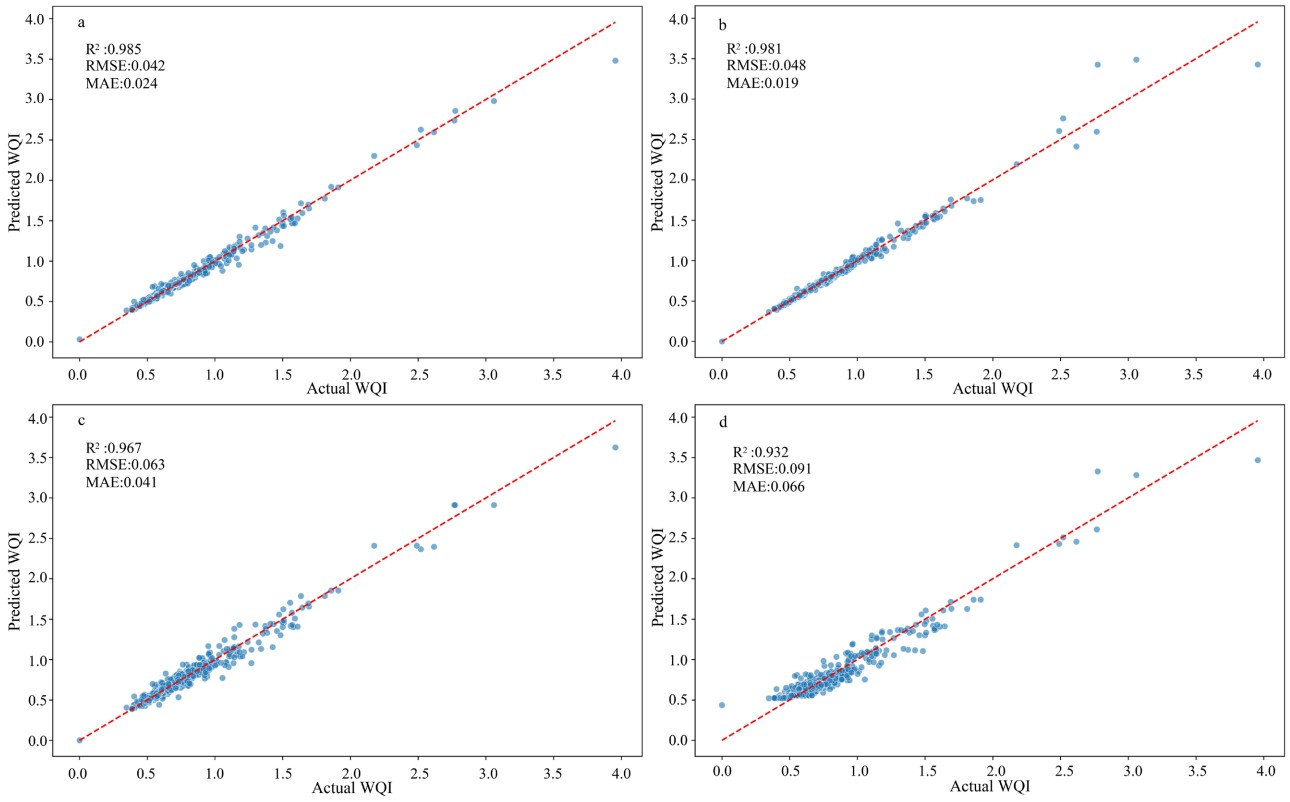

Figure 2. Regression fitting plots of test-set predictions for four machine learning models (the red dashed line represents y = x)
图 2. 4 种机器学习模型测试集回归预测拟合图(红色虚线代表 y = x)

Table 2. Evaluation of prediction performance of four machine learning models
表 2. 4 种机器学习模型预测性能评估

模型类别	R^2	RMSE	MAE
随机森林	0.985	0.042	0.024
极端梯度提升	0.981	0.048	0.019
决策树	0.967	0.063	0.041
自适应增强	0.932	0.041	0.066

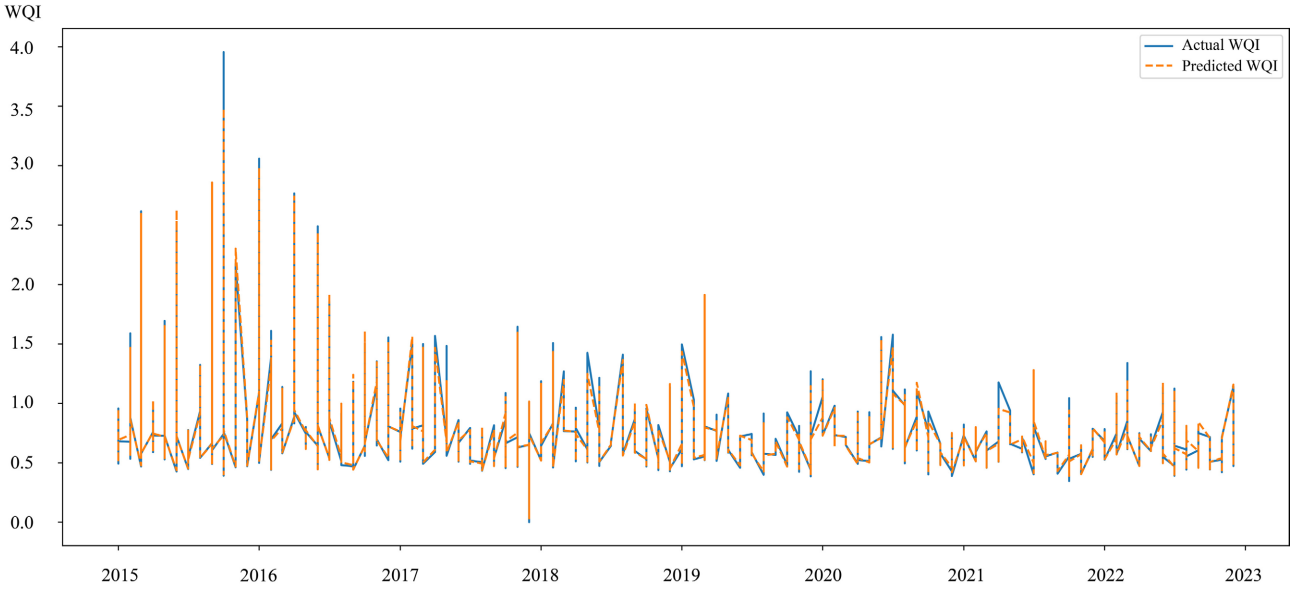

Figure 3. Comparison between observed and predicted water quality values using the Random Forest model
图 3. 随机森林模型水质预测实际值与预测值对比

基于 RF 模型估算的观测值与预测值水质指数(WQI)的时间动态特征如时间序列图所示，时间范围覆盖 2015~2022 年。WQI 作为衡量水质退化程度的综合指标，其数值越高代表水质状况越差。整体来看，数据呈现出明显的两阶段演变特征：在初始阶段(2015~2016 年)，实际值与预测值的 WQI 均表现出显著的时间波动，观测 WQI 一度接近 4.0，表明研究区在该时期经历了严重的水体污染事件。RF 模型能够较好地捕捉这一阶段的高频波动特征及整体上升趋势，但对极端 WQI 峰值的幅度存在一定低估，这也是基于树模型的集成算法在极端事件样本有限条件下进行外推预测时的常见局限。

自 2017 年起，系统进入相对稳定阶段，观测值与预测值的 WQI 均在较窄范围内波动(约 0.5~1.5)，且波动幅度明显减小。WQI 波动性的降低表明水质稳定性得到改善，这一变化可能与研究区污染控制措施的加强及生态修复工程的持续推进密切相关。在整个研究时段内，RF 模型在刻画长期变化趋势和中等幅度波动方面表现出较强的预测能力，预测结果与观测值在峰值和谷值出现的时间上具有良好一致性，说明模型能够有效学习并重现水质变化的潜在动力学特征。然而，在部分细尺度高频波动阶段(如 2019 年和 2021 年的局部偏差)以及对极端污染事件的低估问题仍然存在，这表明模型仍有进一步优化空间，例如引入异常检测算法，或通过增加极端事件样本以增强模型对稀有污染事件的泛化能力。总体而言，本研究结果表明 RF 模型在水质波动幅度适中的情景下具有较高的预测可靠性，同时也强调了针对极端事件预测开展针对性改进的必要性。2016 年之后 WQI 波动性的显著降低进一步为区域水环境治理成效提供了经验证据，并可为未来水资源保护与管理政策的制定提供科学依据。

3.2. 模型局限性与未来研究方向

尽管 RF 模型在整体预测性能方面表现稳健，但其在极值预测上的不足仍需加以说明。该局限性主要源于 RF 的集成平均机制，即通过对多棵决策树预测结果取平均来提高模型稳定性，这在一定程度上会平滑预测结果，从而削弱模型对分布尾部极端高值或低值的敏感性。对于水质指标这类具有明显偏态分布和突发性极端事件特征的环境数据而言，该问题尤为值得关注。

针对这一不足，未来研究可考虑引入更具针对性的改进策略。例如，分位数回归森林(Quantile Regression Forests, QRF)作为 RF 的扩展形式，能够估计条件分布的不同分位数，在刻画水质变量的上下极值方面具有潜在优势。此外，对目标变量进行对数变换或 Box-Cox 变换可有效缓解数据偏态性和异方差问题，从而提高模型对高浓度事件的响应能力。综合采用上述方法，有望在保持树模型稳健性和可解释性的同时，进一步提升对水质极端变化的预测能力。

4. 结论

对四种机器学习模型——RF、XGBoost、DT 和 AdaBoost——的性能评估结果表明，各模型在预测精度上存在明显差异(表 1)。其中，RF 模型表现最优，其 R^2、RMSE 和 MAE 分别为 0.985、0.042 和 0.024，显示出极高的预测精度与稳定性。XGBoost 模型次之，R^2 为 0.981，RMSE 为 0.048，MAE 为 0.019，具有较强的预测能力，但在一致性上略低于 RF 模型。DT 模型($R^2 = 0.967$)的性能相对中等，表明其单树结构在捕捉复杂非线性关系方面存在一定局限。相比之下，AdaBoost 模型表现最弱($R^2 = 0.932$)，可能由于其对噪声较为敏感，在具有异质性的环境数据中容易出现过拟合。总体而言，集成学习模型(RF 与 XGBoost)的表现显著优于单一树模型和基于迭代提升的算法，表明其在处理非线性关系和抑制过拟合方面具有更强的优势。因此，以后的研究可以将 RF 模型作为水质空间预测分析的最优模型，以开展更为精确和系统的水质动态模拟。

参考文献

[1] 汪心雯, 刘子琦, 郭琼琼, 等. 贵州黄洲河流域水质时空分布特征及污染源解析[J]. 环境工程, 2021, 39(9): 69-75.

[2] 马克明, 孔红梅, 关文彬, 等. 生态系统健康评价: 方法与方向[J]. 生态学报, 2001, 21(12): 2106-2116.

[3] 杨丽蓉, 陈利顶, 孙然好. 河道生态系统特征及其自净化能力研究现状与发展[J]. 生态学报, 2009, 29(9): 5066-5075.

[4] 高雯媛, 邹霖, 朱俊毅, 等. 湖南省地表水水质时空变化特征及驱动因子分析[J]. 环境工程, 2024, 42(8): 17-24.

[5] 雷川华, 吴运卿. 我国水资源现状、问题与对策研究[J]. 节水灌溉, 2007(4): 41-43.

[6] Scanlon, B.R., Jolly, I., Sophocleous, M. and Zhang, L. (2007) Global Impacts of Conversions from Natural to Agricultural Ecosystems on Water Resources: Quantity versus Quality. *Water Resources Research*, **43**, W03437. https://doi.org/10.1029/2006wr005486

[7] 张薇, 赵亚娟. 国际水资源现状与研究热点[J]. 地质通报, 2009, 28(2): 177-183.

[8] 李慧. 全球水资源未来可持续性研究[J]. 水利水电快报, 2023, 44(3): 5.

[9] 周佳君. 水资源现状及保护应对措施分析[J]. 城市建设理论研究(电子版), 2020(8): 44.

[10] Zhi, W., Appling, A.P., Golden, H.E., Podgorski, J. and Li, L. (2024) Deep Learning for Water Quality. *Nature Water*, **2**, 228-241. https://doi.org/10.1038/s44221-024-00202-z

[11] Liang, Y., Ke, S., Zhang, J., Yi, X. and Zheng, Y. (2018) GeoMAN: Multi-Level Attention Networks for Geo-Sensory Time Series Prediction. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, Stockholm, 13-19 July 2018, 3428-3434. https://doi.org/10.24963/ijcai.2018/476

[12] Liu, Y., Zhang, Q., Song, L. and Chen, Y. (2019) Attention-Based Recurrent Neural Networks for Accurate Short-Term and Long-Term Dissolved Oxygen Prediction. *Computers and Electronics in Agriculture*, **165**, Article ID: 104964. https://doi.org/10.1016/j.compag.2019.104964

[13] Zhi, W., Ouyang, W., Shen, C. and Li, L. (2023) Temperature Outweighs Light and Flow as the Predominant Driver of Dissolved Oxygen in US Rivers. *Nature Water*, **1**, 249-260. https://doi.org/10.1038/s44221-023-00038-z

[14] Blaen, P.J., Khamis, K., Lloyd, C.E.M., Bradley, C., Hannah, D. and Krause, S. (2016) Real-Time Monitoring of Nutrients and Dissolved Organic Matter in Rivers: Capturing Event Dynamics, Technological Opportunities and Future Directions. *Science of the Total Environment*, **569**, 647-660. https://doi.org/10.1016/j.scitotenv.2016.06.116

[15] Ebeling, P., Kumar, R., Weber, M., Knoll, L., Fleckenstein, J.H. and Musolff, A. (2021) Archetypes and Controls of Riverine Nutrient Export across German Catchments. *Water Resources Research*, **57**, e2020WR028134. https://doi.org/10.1029/2020wr028134

[16] Creed, I.F., Lane, C.R., Serran, J.N., Alexander, L.C., Basu, N.B., Calhoun, A.J.K., *et al*. (2017) Enhancing Protection for Vulnerable Waters. *Nature Geoscience*, **10**, 809-815. https://doi.org/10.1038/ngeo3041

[17] 姚亚. 数据预处理和直方图时间序列在水质预测中的应用[D]: [硕士学位论文]. 杭州: 浙江大学, 2013.

[18] Li, F., Li, D., Wei, Y., Ma, D. and Ding, Q. (2010) Dissolved Oxygen Prediction in *Apostichopus japonicus* Aquaculture Ponds by BP Neural Network and AR Model. *Sensor Letters*, **8**, 95-101. https://doi.org/10.1166/sl.2010.1208

[19] Zhou, J., Wang, Y., Xiao, F., Wang, Y. and Sun, L. (2018) Water Quality Prediction Method Based on IGRA and LSTM. *Water*, **10**, Article No. 1148. https://doi.org/10.3390/w10091148

[20] Zhou, S., Song, C., Zhang, J., Chang, W., Hou, W. and Yang, L. (2022) A Hybrid Prediction Framework for Water Quality with Integrated W-ARIMA-GRU and LightGBM Methods. *Water*, **14**, Article No. 1322. https://doi.org/10.3390/w14091322

[21] Adnan, R.M., Liang, Z., Heddam, S., Zounemat-Kermani, M., Kisi, O. and Li, B. (2020) Least Square Support Vector Machine and Multivariate Adaptive Regression Splines for Streamflow Prediction in Mountainous Basin Using Hydro-Meteorological Data as Inputs. *Journal of Hydrology*, **586**, Article No. 124371. https://doi.org/10.1016/j.jhydrol.2019.124371

[22] Jordan, M.I. and Mitchell, T.M. (2015) Machine Learning: Trends, Perspectives, and Prospects. *Science*, **349**, 255-260. https://doi.org/10.1126/science.aaa8415

[23] Seifeddine, M., Bradai, A., Bukhari, S.H.R., *et al*. (2020) A Survey on Machine Learning in Internet of Things: Algorithms, Strategies, and Applications. *Internet of Things*, **12**, Article ID: 100314.

[24] Akhtar, N., Ishak, M.I.S., Ahmad, M.I., Umar, K., Md Yusuff, M.S., Anees, M.T., *et al*. (2021) Modification of the Water Quality Index (WQI) Process for Simple Calculation Using the Multi-Criteria Decision-Making (MCDM) Method: A Review. *Water*, **13**, Article No. 905. https://doi.org/10.3390/w13070905

[25] Patel, D.D., Mehta, D.J., Azamathulla, H.M., Shaikh, M.M., Jha, S. and Rathnayake, U. (2023) Application of the Weighted Arithmetic Water Quality Index in Assessing Groundwater Quality: A Case Study of the South Gujarat Region. *Water*, **15**, Article No. 3512. https://doi.org/10.3390/w15193512

[26] Uddin, M.G., Nash, S., Rahman, A. and Olbert, A.I. (2022) A Comprehensive Method for Improvement of Water Quality Index (WQI) Models for Coastal Water Quality Assessment. *Water Research*, **219**, Article ID: 118532. https://doi.org/10.1016/j.watres.2022.118532

[27] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[28] Jiang, F., Shi, X., Shi, F., Jia, Z., Song, X., Pu, T., *et al*. (2025) Scale-Dependent Drivers of Water Use Efficiency across China: Integrating Stable Isotopes, Remote Sensing, and Machine Learning. *Catena*, **260**, Article ID: 109403. https://doi.org/10.1016/j.catena.2025.109403

[29] Chen, L., Zhou, J., Guo, L., Bian, X., Xu, Z., Chen, Q., *et al*. (2024) Global Distribution of Mercury in Foliage Predicted by Machine Learning. *Environmental Science & Technology*, **58**, 15629-15637. https://doi.org/10.1021/acs.est.4c00636

[30] Hu, J. and Szymczak, S. (2023) A Review on Longitudinal Data Analysis with Random Forest. *Briefings in Bioinformatics*, **24**, bbad002. https://doi.org/10.1093/bib/bbad002

[31] Min, C., Liao, G., Wen, G., *et al*. (2023) Ensemble Interpretation: A Unified Method for Interpretable Machine Learning.

[32] Hajihosseinlou, M., Maghsoudi, A. and Ghezelbash, R. (2023) A Novel Scheme for Mapping of MVT-Type Pb-Zn Prospectivity: LightGBM, a Highly Efficient Gradient Boosting Decision Tree Machine Learning Algorithm. *Natural Resources Research*, **32**, 2417-2438. https://doi.org/10.1007/s11053-023-10249-6

[33] Zhou, Z.-H. (2025) Ensemble Methods: Foundations and Algorithms. CRC Press. https://doi.org/10.1201/9781003587774

[34] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., *et al*. (2020) From Local Explanations to Global Understanding with Explainable AI for Trees. *Nature Machine Intelligence*, **2**, 56-67. https://doi.org/10.1038/s42256-019-0138-9

[35] Natekin, A. and Knoll, A. (2013) Gradient Boosting Machines, a Tutorial. *Frontiers in Neurorobotics*, **7**, Article No. 21.

https://doi.org/10.3389/fnbot.2013.00021

[36] Bian, W., Fang, J., Wang, P., Sun, Q., Fang, J., Kong, F., *et al*. (2025) Deep Learning Surrogate Models for Spatiotemporal Prediction of Coastal Flooding Inundations in Tianjin, China. *Journal of Hydrology*: *Regional Studies*, **60**, Article ID: 102593. https://doi.org/10.1016/j.ejrh.2025.102593

[37] Xue, Y., Liang, H., Zhang, B. and He, C. (2022) Vegetation Restoration Dominated the Variation of Water Use Efficiency in China. *Journal of Hydrology*, **612**, Article ID: 128257. https://doi.org/10.1016/j.jhydrol.2022.128257

[38] Liang, Y., Ding, F., Liu, L., Yin, F., Hao, M., Kang, T., *et al*. (2025) Monitoring Water Quality Parameters in Urban Rivers Using Multi-Source Data and Machine Learning Approach. *Journal of Hydrology*, **648**, Article ID: 132394. https://doi.org/10.1016/j.jhydrol.2024.132394