J. Lake Sci.(湖泊科学),2025,**37**(3):915-927
DOI 10.18307/2025.0328

基于机器学习的鄱阳湖溶解氧波动特征及预测*

李晓瑛[1,2],王　华[1,2**],吴小毛[3],吴　怡[1,2],徐浩森[1,2]
(1:河海大学环境学院,南京 210098)
(2:河海大学浅水湖泊综合治理与资源开发教育部重点实验室,南京 210098)
(3:江西省鄱阳湖水利枢纽建设办公室,南昌 330009)

摘　要:溶解氧(DO)作为反映水体自净能力和水环境质量的关键指标,是评估鄱阳湖水体健康状况的重要参数。随机森林(RF)和改进支持向量回归(PSO-SVR)2 种机器学习的高效算法被引入到鄱阳湖 DO 的预测工作中,时间上选择 1988－2023 年水质数据进行预测,空间上挑选了位于鄱阳湖和入湖 5 条河流的共 8 个关键监测站点:棠荫、信江东支、鄱阳、赣江主支、抚河口、修河口、康山和湖口。对 8 个监测站点的 DO 进行曼肯达尔趋势检验,整体上 DO 浓度上升的站点为抚河口、修河口、康山和湖口,其中康山和湖口的 DO 浓度在后期表现出显著上升趋势。基于随机森林重要性指数(IMI)探究了 DO 与其他水质因子间的响应关系,在 8 个监测站点中水温(T)对 DO 的重要性指数均较高,其次是高锰酸盐指数(COD_{Mn}),各个因子的平均 IMI 排序为 T>COD_{Mn}>TN>NH_3-N>TP>pH,其重要性指数值分别为 2.54、0.81、0.65、0.63、0.43 和 0.37。使用 RF 和 PSO-SVR 模型对 1988－2023 年月均水质数据进行预测对比分析。整体上,RF 和 PSO-SVR 模型在 8 个监测站点的总体平均误差分别为 0.32 和 0.54。基于混淆矩阵的模型性能评价中,RF 和 PSO-SVR 模型的平均准确率 η 分别为 0.67 和 0.52。模型在训练集上整体预测性能为:RF(R^2=0.953;*RMSE*=0.397 mg/L)>PSO-SVR(R^2=0.822;*RMSE*=0.764 mg/L)。模型在预测集上整体预测性能为:RF(R^2=0.836;*RMSE*=0.660 mg/L)>PSO-SVR(R^2=0.815;*RMSE*=0.686 mg/L)。两种模型均表现出优秀的预测性能,其中 RF 的预测能力更好。引入机器学习的高效算法实现对鄱阳湖 DO 进行精准预测,以期揭示鄱阳湖水质规律以及水质因子之间的内在联系,为环境监测与管理提供科学的决策支持。
关键词:鄱阳湖;溶解氧;预测;随机森林;支持向量回归;混淆矩阵

Characterization and prediction of dissolved oxygen fluctuation in Lake Poyang based on machine learning*

Li Xiaoying[1,2], Wang Hua[1,2**], Wu Xiaomao[3], Wu Yi[1,2] & Xu Haosen[1,2]
(1: *College of Environment, Hohai University, Nanjing* 210098, *P.R.China*)
(2: *Key Laboratory of Integrated Regulation and Resource Development on Shallow Lake of Ministry of Education, Hohai University, Nanjing* 210098, *P.R.China*)
(3: *Jiangxi Province Poyang Lake Water Conservancy Center Construction Office, Nanchang* 330009, *P.R.China*)

Abstract: Dissolved oxygen (DO) is a key indicator reflecting the self-purification ability of water bodies and the quality of water environment. DO is also an important parameter for assessing the health of water bodies in Lake Poyang. In this study, two efficient machine learning algorithms, random forest (RF) and improved support vector regression (PSO-SVR), were introduced into the monitoring and prediction of DO in Lake Poyang. The water quality data from 1988 to 2023 were selected for prediction in time, and a total of eight key monitoring stations of Lake Poyang and five rivers entering the lake were spatially selected: Tangyin, east branch of Xinjiang River, Poyang, main branch of Ganjiang River, Fuhekou, Xiuhekou, Kangshan and Hukou. Firstly, Mann-Kendall trend test was performed on the DO of the eight monitoring stations. The stations with overall increasing DO were Fuhekou, Xiuhekou, Kangshan and Hukou, among which Kangshan and Hukou showed a significant increasing trend in the later stage. Sec-

* 2024-04-21 收稿;2024-09-02 收修改稿。
国家重点研发计划项目(2023YFC3209000001)和江西省“科技+水利”联合计划项目(2023KSG003)联合资助。
** 通信作者;E-mail: wanghua@hhu.edu.cn。

ondly, the response and relationship between DO and other water quality factors were explored based on the random forest importance index (IMI). The importance index of water temperature (T) to DO was higher in all 8 monitoring stations, followed by month, and the average IMI of each factor ranked T>COD_{Mn}>TN> NH_3-N>TP>pH, with importance index values of 2.54, 0.81, 0.65, 0.63, 0.43 and 0.37, respectively. The model predictions were then analyzed in comparison to the monthly average water quality data from 1988 to 2023 using RF and PSO-SVR. Overall, the overall mean errors were 0.32 for the RF model and 0.54 for the PSO-SVR model at the eight monitoring stations. The mean accuracies η *i*n the model performance evaluation based on the confusion matrix were 0.67 for RF and 0.52 for PSO-SVR, respectively. The overall prediction performances on the training set were RF (R^2 = 0.953; *RMSE* = 0.397 mg/L)>SVR (R^2 = 0.822; *RMSE* = 0.764 mg/L). The overall prediction performance of the models on the prediction set was RF (R^2 = 0.836; *RMSE* = 0.660 mg/L)>SVR (R^2 = 0.815; *RMSE* = 0.686 mg/L). Both models showed excellent predictive performance, with RF having better predictive ability. The R^2 values of the RF model were more concentrated in the training and prediction sets, indicating that the model had better stability and generalization ability. The *RMSE* values were also more concentrated in the training and prediction sets, but slightly higher in the prediction set. The R^2 and *RMSE* values of the PSO-SVR model were more dispersed in the training and test sets, indicating that the model's performance varied greatly in different cross-sections, and it may need to be adjusted for different data characteristics. Overall, the RF model showed the best prediction ability on all monitoring sections, with the highest R^2 value and the lowest *RMSE* value, and showed excellent performance and generalization ability on both training and test sets. The PSO-SVR model also performed well on most monitoring sections, and its prediction performance was slightly inferior to that of the RF model, and it may need to optimize the structure or parameters of the model to improve the prediction accuracy and stability. improve the prediction accuracy and stability. Both models showed excellent predictive performance, with RF having better predictive ability. An efficient algorithm of machine learning was introduced to realize the accurate prediction of dissolved oxygen in Lake Poyang, with a view to revealing the water quality pattern of Lake Poyang and the intrinsic connection between the water quality factors, and providing scientific decision support for environmental monitoring and management.

Keywords: Lake Poyang; dissolved oxygen; prediction; random forest; support vector regression; confusion matrix

鄱阳湖作为中国第一大淡水湖,承担着生态平衡、水资源调节和区域经济发展的多重重要角色[1-3]。近年来,随着工农业活动的增加及气候变化的影响,鄱阳湖水环境面临着巨大的压力[4-6]。水质的波动不仅影响着湖泊生态系统的稳定性和生物多样性,也直接关系到周边居民的饮用水安全和地区的可持续发展[7-9]。因此对鄱阳湖水质变化特征的深入研究和准确预测显得尤为重要。鄱阳湖作为通江湖泊与长江相连,特殊的地理位置及复杂的江湖相互关系导致鄱阳湖形成独特而完整的江湖复合生态系统[10-12]。受诸多环境要素,例如湖泊自然水动力条件[13-15]、人为活动[16-18]等影响,鄱阳湖水质呈现出多变性与复杂性。溶解氧(DO)作为反映水体自净能力和水环境质量的关键指标,是评估水体健康状况的重要参数[19-21]。然而,受自然条件和人为活动的双重影响,鄱阳湖 DO 浓度的时空分布呈现出复杂的动态变化,这对于鄱阳湖水质监测和预警提出了更高的要求。

曼肯达尔(Mann-Kendall,MK)趋势检验法是一种气候变化及预测评估方法,该方法通常用于检测时间序列数据中是否存在显著的趋势,是气候气象研究和水质预测领域的常用工具。如姚嘉伟等[22]以中国 6 个市级水源地为研究对象,采用水质指数法结合 MK 趋势检验法,深入分析了各水源地的水质突变时间及演变趋势。DO 浓度具有非线性、时序性和不稳定性等特点,增加了准确预测的难度,使用 MK 趋势检验法进行 DO 浓度变化特征分析对于后续预测具有一定的指导意义。Chi 等[23]提出一种基于 WT-MIC-GRU 方法的鄱阳湖 DO 浓度预测模型,为湖泊水体及湖泊水质监测数据中缺失值的修复提供参考。基于统计学习理论的随机森林(random forest, RF)是近几年迅速发展起来的一种分类和预测模型,为水质评价研究提供了一条新的途径,被广泛应用于水质模型的预测和优化过程中[24-26],Geetha 对印度 Chittar Pattanam Channel、甘尼亚古马里县、泰米尔纳德邦的水质参数进行预测,结果表明 RF 算法在预测精度方面优于其他模型,平均绝对误差为 0.56,均方误差(*MAE*)为 0.33,均方根误差(*RMSE*)为 0.56[27]。支持向量回归模型(SVR)特别适用于处理非线性和高维数据的情况,使其成为预测水质变化的理想工具。如 Jamal 等使用 SVM 模拟伊朗阿吉柴河水质,证明了 SVM 在所有模拟站点显示出较高的决定系数(R^2)和较低的 *RMSE* 和 *MAE*[28]。尽管鄱阳湖与上述水体具体气候类型不同,但都受到季风的影响,具有明显的湿季和干季之分;农业活动占主导地

位;浅水特性使水质易受外界影响,DO 浓度变化显著,使用 RF 和 SVR 进行鄱阳湖 DO 预测具有适用性和重要研究价值。

本研究基于 1988－2023 年鄱阳湖及其入湖 5 条主要河流的实测资料,深入分析了水温(T)、pH、高锰酸盐指数(COD_{Mn})、氨氮(NH_3-N)、总氮(TN)、总磷(TP)等水质参数的变化特征,并特别关注 DO 的时空动态。选取了棠荫、信江东支、鄱阳、赣江主支、抚河口、修河口、康山和湖口 8 个关键监测站点的水质数据,采用改进的支持向量机回归(PSO-SVR)和 RF 机器学习模型,对历年的月均水质数据进行了预测和对比分析,以期揭示水质因子之间的内在联系,为环境监测与管理提供科学的决策支持。

1 材料与方法

1.1 研究区域

鄱阳湖(28°25′~29°45′N,115°50′~116°44′E)位于江西省北部,长江中下游南岸,区域位置如图 1 所示。鄱阳湖平均水深 8.4 m,最深处 25.1 m 左右,总蓄水量约为 276 亿 m^3,是中国最大的淡水湖泊。上游承接赣江、抚河、信河、饶河、修河 5 条主要河流来水,成为全省的“集水盆”、“五河”入江的“中转站”[29-34]。鄱阳湖水质及水动力流场受“五河七口”来水影响,同时作为通江湖泊与长江相连,特殊的河湖交互关系导致其水质现状在时间和空间维度均呈现出异质性。

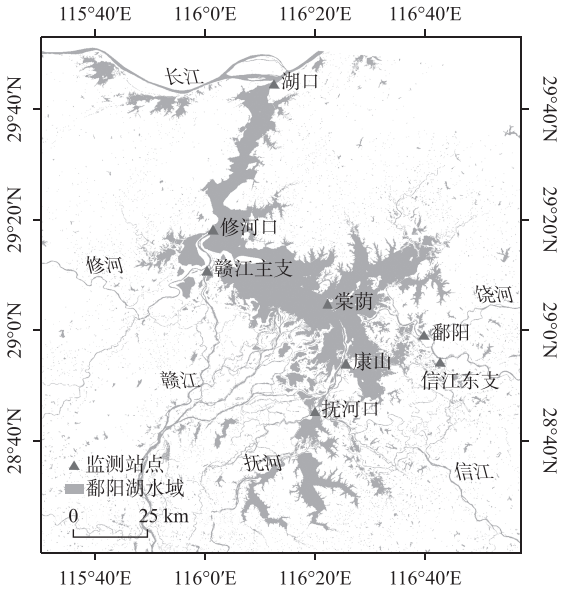

图 1 鄱阳湖区域位置及监测站点示意图

Fig.1 Schematic diagram of the location and monitoring stations of Lake Poyang

1.2 数据收集与处理

选择棠荫、信江东支、鄱阳、赣江主支、抚河口、修河口、康山和湖口 8 个鄱阳湖代表性水质监测站点,选取的输入变量为 T、pH、NH_3-N、TN、TP、COD_{Mn} 6 个水质因子,选取的目标变量为 DO,对各个监测点位 1988－2023 年的逐月水质监测数据进行预测分析。为保证多项水质因子的数据完整性,其中 2018－2020 年数据不包含在内。数据来源为江西省水文局和江西省水文监测中心(官方网站 http://www.jxssw.gov.cn/)。监测资料执行《水环境监测规范》(SL219－2013)及《国家地表水水质自动监测站补充监测技术规定》(2017 年修订版)等规定,超标限根据《地表水环境质量标准》(GB 3838－2002)Ⅲ类标准执行。

1.3 MK 趋势检验法

MK 趋势检验法是一种气候变化及预测评估方法,常被用在气候气象、水质预测等方面[35-37],其研究计算步骤为:

对于样本个数为 n 的时间序列,Mann-Kendall test 统计先计算 S_k:

$$S_k = \sum_{i=1}^{k} r_i, k = 2,3,\cdots,n \tag{1}$$

MK 趋势检验分析早期和后期数据点之间的符号差异,式(1)中 r_i 的取值为:

$$r_i = \begin{cases} 1, x_i > x_j \\ 0, x_i \leqslant x_j \end{cases}, j = 1,2,\cdots,i \tag{2}$$

可见秩序列 S_k 是第 i 时刻数值大于 j 时刻数值个数的累计数。假设时间序列是随机独立的,对统计量进行定义:

$$UK_k = \frac{|S_k - E(S_k)|}{\sqrt{\mathrm{var}(S_k)}}, k = 1,2,\cdots,n \tag{3}$$

式中,$UK_{k1} = 0$,$E(S_k)$ 和 $\mathrm{var}(S_k)$ 是 S_k 的均值和方差。在时间序列数据相互独立,且具有相同连续分布时,它们可由下式算出:

$$E(S_k) = \frac{k(k-1)}{4} \tag{4}$$

$$\mathrm{var}(S_k) = \frac{k(k-1)(2k+5)}{72} \tag{5}$$

MK 检验统计量结果主要通过 UF_k 值进行分析,以±1.96 为分界标准。如果 $UF_k>0$,说明该序列在描述时间段内处于上升的趋势;反之则说明该序列处于下降的趋势。如果 $|UF_k|>1.96$,则说明该序列具有明显的上升或下降趋势。UB_k 统计量是基于顺序统计量计算的,用于评估时间序列数据中某一点之前所有数据点的趋势。如果 UF_k 和 UB_k 2 条曲线在临界线之间出现交点,则交点对应的时刻即为突变开始的时间;若交点出现在临界线外或出现多个交点,可结合其他检验方法进一步判定是否为突变点。

1.4 改进支持向量机回归预测模型 PSO-SVR

支持向量机(SVM)是一种基于结构风险最小化原理的强大机器学习算法,专注于数据分类和回归分析。支持向量回归(SVR)是 SVM 的一个应用,用于连续数据的预测和拟合。实施 SVR 核心任务之一是确定模型的关键参数,如核函数的 γ 值和惩罚系数 α。改进的 PSO-SVR 模型结合了粒子群优化(PSO)算法和支持向量回归(SVR),以优化 SVR 的关键参数,提高模型在复杂数据集上的性能。PSO 通过迭代逐步搜索全局最优解,结合位置和速度更新就可以实现数据空间的最优解搜索。鄱阳湖的 DO 浓度受到水环境中多因素相互作用的影响,在这样的环境中,改进后的 PSO-SVR 机器学习模型可以胜过传统的预测模型。在 PSO 中,粒子通过跟踪个体历史最优位置(个体最优,*pBest*)和群体历史最优位置(全局最优,*gBest*)来更新自己的位置和速度。粒子的速度和位置更新公式如下:

$$v_i^{(t+1)} = \omega v_i^{(t)} + c_1 r_1 (pBest_i - x_i^{(t)}) + c_2 r_2 (gBest - x_i^{(t)}) \tag{6}$$

$$x_i^{(t+1)} = x_i^{(t)} + v_i^{(t+1)} \tag{7}$$

式中,$v_i^{(t)}$ 是粒子 i 在时间 t 的速度,$x_i^{(t)}$ 是粒子 i 在时间 t 的位置,ω 是惯性权重,控制前一速度对当前速度的影响,c_1 和 c_2 是学习因子,r_1 和 r_2 是[0,1]区间内的随机数,*pBest* 是粒子 i 目前找到的最优位置,*gBest* 是所有粒子中找到的最优位置。通过这种方式,PSO 帮助找到最优的 α 和 γ 参数,使得 SVR 模型在特定数据集上达到最佳性能。PSO-SVR 模型特别适用于高维和非线性的数据集,能够有效提升 SVR 模型的预测精度和泛化能力。

1.5 随机森林算法 RF

DO 的年际变化数据存在非线性趋势、周期性和季节性等规律,随机森林在训练过程中基于决策树的集成方法建模非线性关系,可以更好地适应时间序列数据的特点。在预测过程中,水质指标数据构成高维数据空间,RF 算法自动选择重要特征,并在每棵树的节点上进行特征子集的随机选择,从而在预测水质数据时显示出高准确性。RF 模型基本原理是构建多个决策树,并将它们的预测结果进行整合。每棵树都是独立建立的,首先进行自助采样从原始数据集中使用有放回抽样选取样本,形成多个训练集。然后在每个决策树的每个分裂节点,随机选择一部分候选特征,基于这些特征找到最佳的分裂点,计算所有树预测结果的平均值[38-39]。另外,随机森林可以计算自变量对于预测变量的重要性,得到重要性指数(importance index,

IMI)衡量预测指标与作为自变量的水质指标之间的内在关系,其涉及如下计算:

计算第 i 棵树节点 q 的 Gini 系数:

$$\mathrm{Gini}_q^{(i)} = \sum_{c=1}^{|c|}\sum_{c'\neq c} p_{qc}^{(i)} p_{qc'}^{(i)} = 1 - \sum_{c=1}^{|c|} (p_{qc}^{(i)})^2 \tag{8}$$

式中,c 为总分类类别,p_{qc}为节点 q 中类别 c 的出现概率。

计算每个划分的 Gini 系数:

$$\mathrm{Gini}_q^{(\mathrm{split})(i)} = \mathrm{Gini}_{q\mathrm{left}}^{(i)} + \mathrm{Gini}_{q\mathrm{right}}^{(i)} \tag{9}$$

计算节点 q 分支前后 Gini 系数变化量:

$$\mathrm{IMI}_j^{(\mathrm{Gini})} = \sum_{q\in Q} (\mathrm{Gini}_q^{(i)} - \mathrm{Gini}_q^{(\mathrm{split})(i)}) \tag{10}$$

式中,集合 Q 为 X_j在森林中出现的所有节点构成的集合。

1.6 机器学习性能评价

为了对模型的整体预测效能进行定量衡量,基于混淆矩阵结果计算准确率 η。该准确率指标通过以下公式准确定义:它是混淆矩阵主对角线元素之和(即模型正确预测的实例数量)与所有预测值总数的比值。这一比值反映了模型在整个数据集上的预测准确性,即模型正确预测结果所占的比例,从而提供了一种简洁而有效的方式来评估模型性能。通过这一指标,可以直观地了解模型在给定任务中的整体表现。η 的计算公式为:

$$\eta = \frac{\sum_{i=1}^{n} C_i}{C_{\mathrm{sum}}} \tag{11}$$

式中,C_i为混淆矩阵主对角线元素数量,C_{sum}为所有预测值总数。

按照 70%~80%的数据用于训练,20%~30%的数据用于测试的划分标准,进行训练集和测试集的划分,具体划分比例可根据各个监测站点的数据量和多次预测最佳结果进行调整。通过 R^2和 *RMSE* 来评估模型分别在训练集和测试集上的具体性能。R^2是衡量模型解释变量总变异程度的统计量,其范围为 0~1,值越接近 1 表示模型对数据的拟合程度越高。计算公式如下:

$$R^2 = 1 - \frac{\sum (y_i - y_i^*)^2}{\sum (y_i - \bar{y})^2} \tag{12}$$

式中,y_i为实际值,y_i^* 为预测值,$\bar{y}$ 为实际值的平均值。

RMSE 是衡量预测误差的标准偏差,其值越小表示模型的预测误差越小。计算公式如下:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - y_i^*)^2} \tag{13}$$

式中,n 为样本数量。

2 结果与讨论

2.1 基于 MK 检验的鄱阳湖溶解氧变化特征分析

对鄱阳湖 8 个站点的 DO 进行 MK 趋势分析,结果如图 2 所示。关注 MK 趋势检验中的 UF 和 UB 统计量的突变时间点及其在不同站点的空间分布,整体上 DO 浓度上升的站点为抚河口、修河口、康山和湖口,其中康山和湖口的 DO 浓度在后期表现出显著上升趋势。其原因可能是近年来对鄱阳湖及其流域的污染物排放进行了有效控制,水质改善,特别是有机污染物减少,显著提高了 DO 水平。具体来看,抚河口从 2012 年 2 月起 UF 统计量持续大于 0,呈现逐年增加的趋势,在 2015 年 3、4 月和 2017 年 2—5 月的 UF>1.96,DO 表现出显著增加的趋势,在 2021 年 7 月左右出现 UF 和 UB 统计量的交点,说明此时 DO 突变增加。修河口相对抚河口的 DO 增加趋势不明显,UF 统计量一直未超过 1.96。康山站从 2013 年 3 月起 UF 突破 1.96 阈值,呈现显著增加趋势,并在 2021 年 7—11 月出现 2 次增加突变。湖口站 UF 线在整个时间序列中持续上升,并在 2000 年 7 月始终位于 1.96 以上,显示出持续的显著上升趋势,尤其在中后期的上升更为显著。对于棠

荫、鄱阳、赣江主支、信江东支 4 个站点,由于 UF 和 UB 线没有明显超过±1.96 的阈值,因此没有明显的突变点,呈现出周期性波动变化的特点。整体上鄱阳湖的 DO 水平呈现上升趋势,在一定程度上体现出湖泊水质向好趋势,但仍需要深入探究各种因素对水质的具体影响,并持续关注鄱阳湖水体健康。

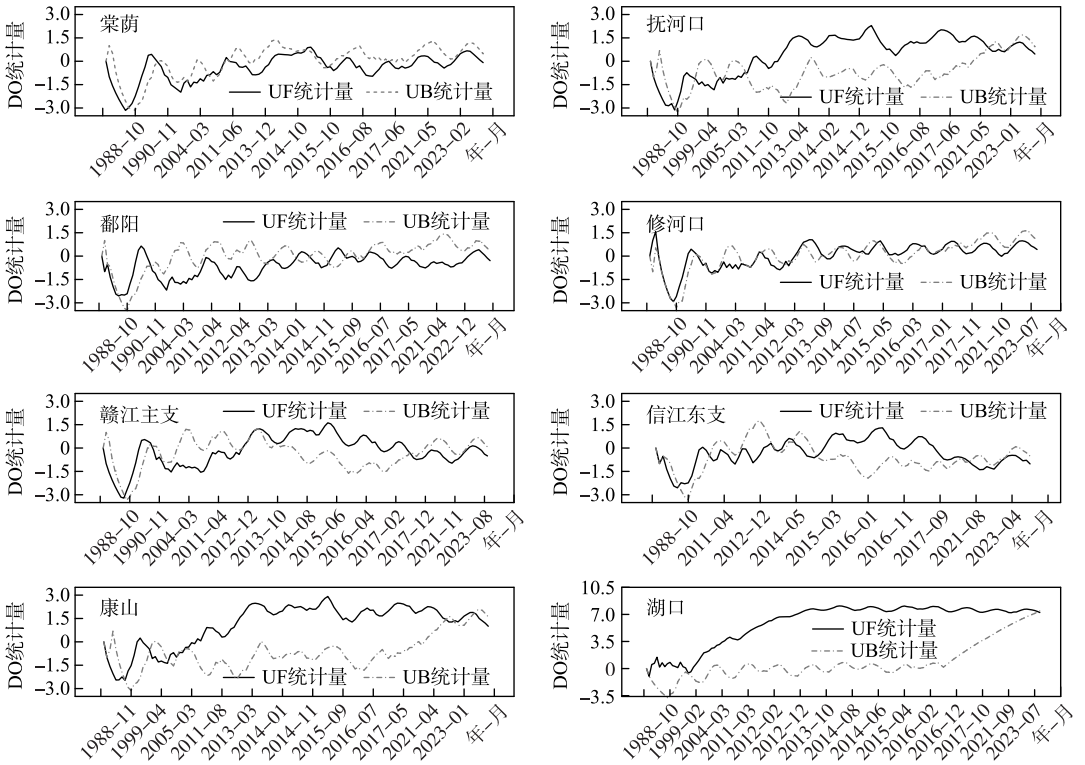

图 2 鄱阳湖 8 个站点 DO 的 MK 趋势分析结果

Fig.2 MK trend analysis results of DO for 8 sites in Lake Poyang

2.2 鄱阳湖水质指标间关系探究

选择棠荫、信江东支、鄱阳、赣江主支、抚河口、修河口、康山和湖口 8 个监测站点,使用随机森林算法进行鄱阳湖 DO 的污染驱动因子重要性指数评价。选择 T、pH、COD_{Mn}、NH_3-N、TP 和 TN 6 个水质因子作为自变量,计算得到每个因子对 DO 的重要性指数 IMI,结果如图 3 所示。

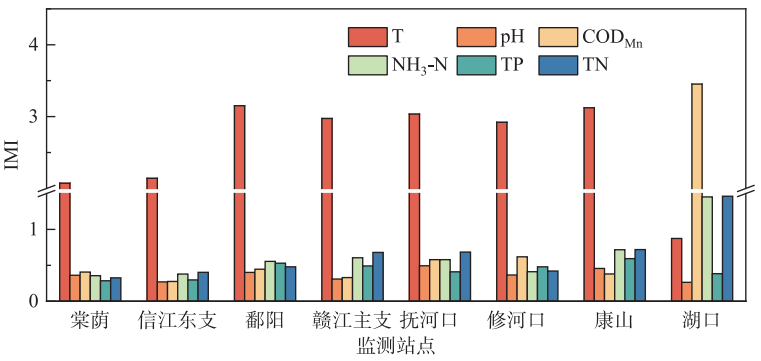

图 3 鄱阳湖 DO 影响因子重要性指数

Fig.3 IMI of influencing factors to DO in Lake Poyang

从结果可以看出,在 8 个监测站点中 T 对 DO 的重要性指数均较高,各个因子的平均 IMI 排名为 T>COD_{Mn}>TN>NH_3-N>TP>pH,其重要性指数值分别为 2.54、0.81、0.65、0.63、0.43 和 0.37。以棠荫站为例,其 DO 驱动因子按照重要性指数从高到低排序依次为 T、COD_{Mn}、pH、NH_3-N、TN 和 TP,T 的 IMI 值为 2.07,远超过其他水质因子,说明 T 的不均匀分布对鄱阳湖 DO 时空分布差异的影响十分显著,这与温度对氧的扩散系数和饱和度的影响密切相关。特别的是湖口站 COD_{Mn} 的 IMI 为 3.45,湖口站位于长江和鄱阳湖的交界处,COD_{Mn} 是一个衡量水体中可被氧化物质消耗氧量的指标,因此它的重要性指出该区域可能面临较高的有机及部分无机污染负荷。TN 和 NH_3-N 两个水质因子在 8 个站点的重要性指数排名均靠前,说明二者对鄱阳湖 DO 浓度变化的影响也较大。氮浓度可能随季节而变化,特别是在农业径流强的春季和秋季,这种季节性的营养盐输入可以导致 DO 水平在一年中的不同时间出现显著变化。

2.3 各监测站点机器学习预测结果分析

选定鄱阳湖 8 个监测站点的水质指标 T、pH、COD_{Mn}、NH_3-N、TN、TP,将以上 6 个影响因素的数值作为输入源,对棠荫、信江东支、鄱阳、赣江主支、抚河口、修河口、康山和湖口 8 个监测站点分别使用 PSO-SVR 和 RF 算法对 DO 进行预测,得到预测值与实际值的对比数据(图 4、5)。

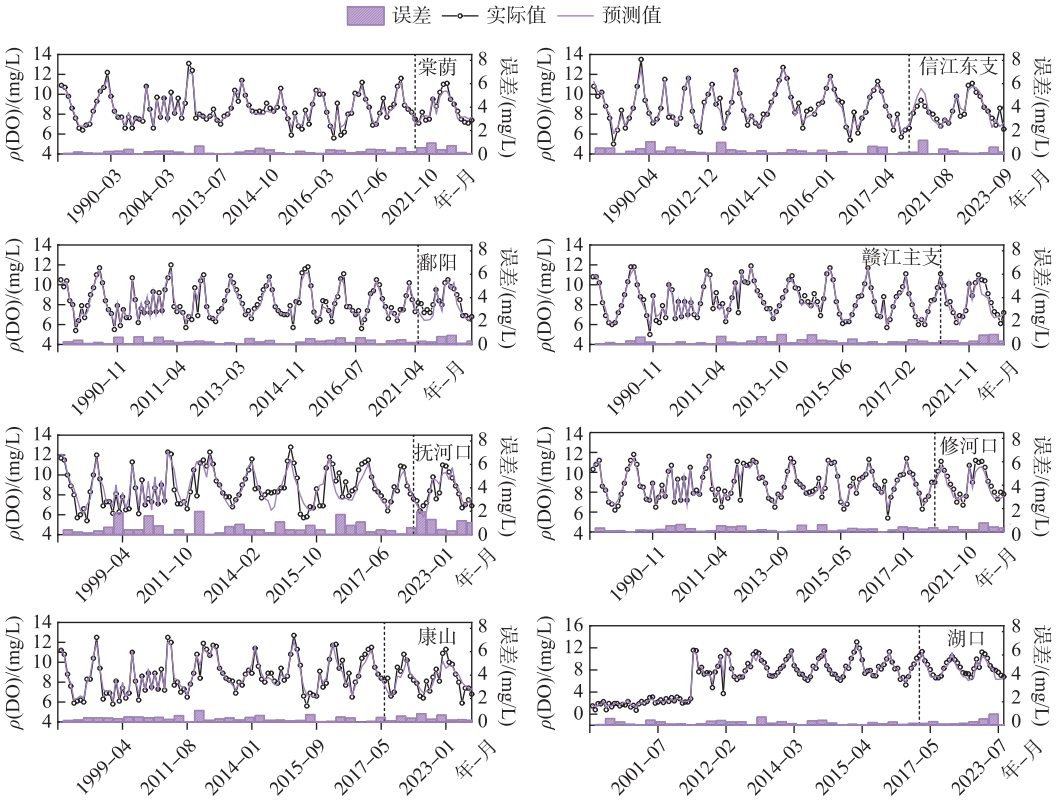

图 4 各监测站点 RF 预测值与实际值对比(虚线前为训练集,虚线后为测试集)

Fig.4 Comparison of predicted and actual values at monitoring stations based on RF

从结果来看,RF 模型(图 4)的预测值与实际值的吻合度较高,尤其是在捕捉数据的周期性波动方面表现突出。各监测站点的误差普遍较低,8 个监测站点的误差均在 0.40 及以下(表 1),表明 RF 模型具有较高的预测准确性。PSO-SVR 模型(图 5)的预测值与实际值非常接近,表明模型也能够较准确地预测 DO 水平。误差整体上较小,说明预测值与实际值之间的差异不大。整体上,RF 和 PSO-SVR 模型在 8 个监测站点的总体平均误差分别为 0.32 和 0.54。在所有站点中,RF 模型预测精度最高的前 3 个站点分别为康山、棠荫和抚

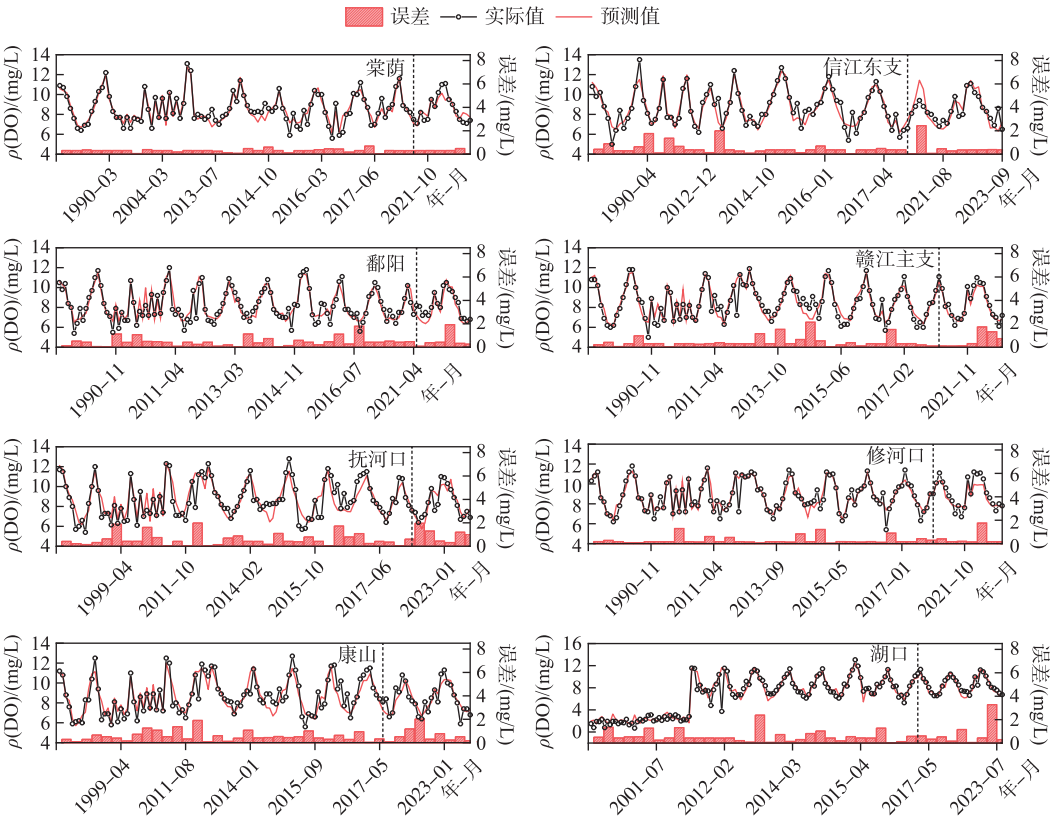

图 5 各监测站点 PSO-SVR 预测值与实际值对比(虚线前为训练集,虚线后为测试集)

Fig.5 Comparison of predicted and actual values at monitoring stations based on PSO-SVR

河口,误差分别为 0.19、0.29 和 0.31;PSO-SVR 模型预测精度最高的前 3 个站点分别为康山、抚河口和棠荫,误差分别为 0.38、0.40 和 0.43。DO 水平呈现显著的周期性波动,这种波动很可能与季节性变量,如温度、水生植物的周期性生长与衰退以及河流的流量变化有关。在夏季,由于水温升高促进了水中氧气的扩散,DO 水平通常会较低;而在冬季,较低的水温和减少的生物活动则导致 DO 水平升高。综合评估显示,尽管使用 PSO-SVR 模型在训练过程中提高了效率,但其预测精度仍低于 RF 模型。

表 1 各监测站点 RF 和 PSO-SVR 模型误差

Tab.1 RF and PSO-SVR model errors at each monitoring station

模型	误差/(mg/L)								
	棠荫	信江东支	鄱阳	赣江主支	抚河口	修河口	康山	湖口	均值
RF	0.29	0.35	0.32	0.35	0.31	0.40	0.19	0.32	0.32
PSO-SVR	0.43	0.58	0.58	0.68	0.40	0.67	0.38	0.60	0.54

2.4 各监测断面机器学习预测模型性能评价

评估预测模型在 DO 这一关键指标上的预测性能时,采用了另一种视角,以混淆矩阵作为主要评估工具来量化与分析 RF 和 PSO-SVR 的模型预测结果,结果如图 6 所示。混淆矩阵提供了一个直观的方式来展示模型预测值与真实值之间的关系,横纵轴分别代表预测值与真实值,并通过矩阵中的元素数量映射来表示特定范围内的预测准确性。每个方块上的数字表示真实的 DO 值被预测成不同值的个数。例如,在棠荫站

的 RF 混淆矩阵结果中,数值 27 代表真实 DO 值在 3 mg/L≤DO<4 mg/L 的范围内且被正确预测的次数为 27;而在抚河口的 RF 混淆矩阵结果中,数值 9 代表真实 DO 值在 5 mg/L≤DO<6 mg/L 范围内但被错误地预测为 4 mg/L≤DO<5 mg/L 的次数为 9,从而表明一种预测偏差。通过混淆矩阵还可以直观判断出预测值偏高或是偏低,在主对角线下方的数值代表预测值比真实值偏高,在主对角线上方的数值代表预测值比真实值偏低。

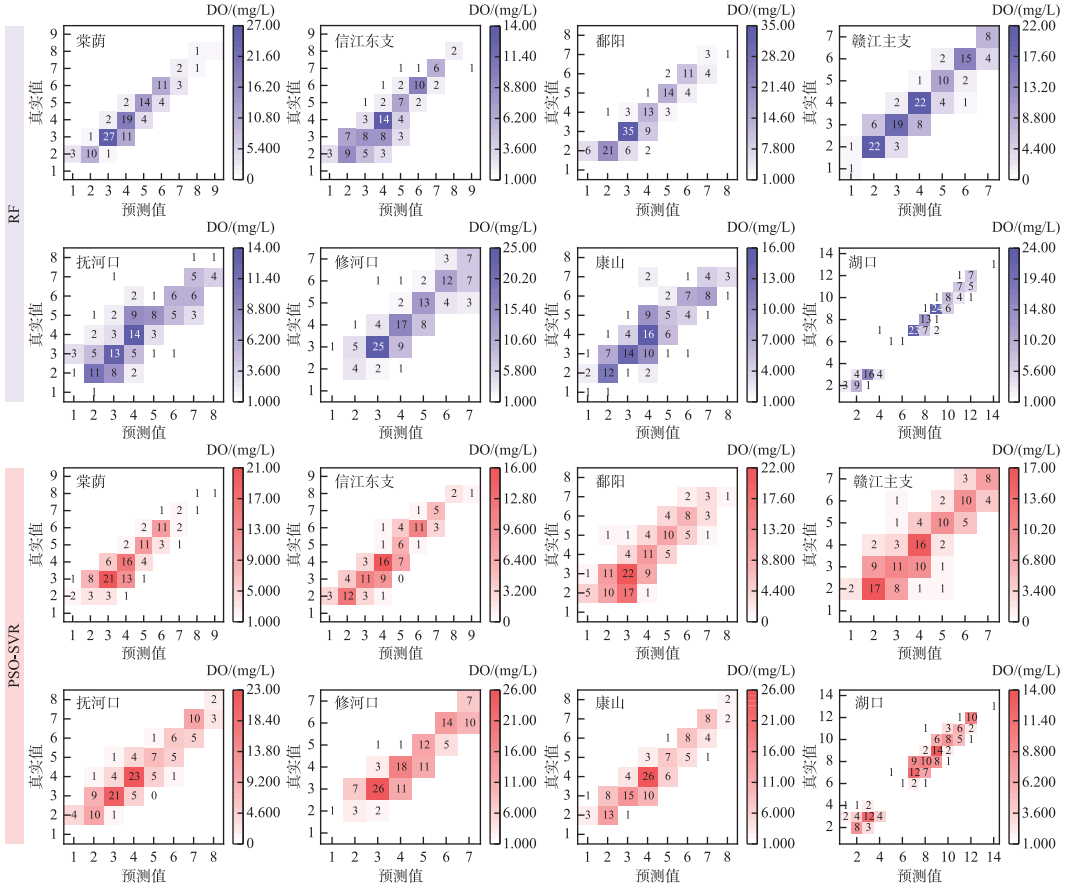

图 6 各监测站点机器学习混淆矩阵

Fig.6 Machine learning confusion matrix for each monitoring station

为了定量衡量模型的预测效能,计算了准确率 η。该指标定义为混淆矩阵主对角线(即正确预测的实例数)之和与所有预测值的总数之比。计算发现 RF 在进行 DO 水平预测时展现出较高的 η。通过在 8 个不同的监测站点上进行评估,该模型取得了较为优异的表现,η 分别为 0.72、0.60、0.70、0.74、0.62、0.61、0.62 和 0.71,其中赣江主支的准确率为最高。与 RF 模型相比,PSO-SVR 的 η 稍低,8 个不同的监测站点分别为 0.56、0.53、0.46、0.55、0.45、0.59、0.46 和 0.53。PSO-SVR 在抚河口的预测任务中结果存在着普遍偏高的趋势。这种偏差有可能源于模型结构、训练数据的质量或是预测环境的特殊性质等多种因素。RF 和 PSO-SVR 模型的平均 η 分别为 0.67 和 0.52。

综合来看 RF 模型的性能优于 PSO-SVR 模型,尽管使用 PSO 算法优化了 SVR,提高了其预测效率,但预测精确度仍不如 RF 模型。RF 模型以其能够建模非线性关系和自动筛选重要特征,在高维数据空间中预测水质表现出了较高的准确性。因此,RF 模型在所有监测断面上均表现出最佳预测能力,具有最高的 R^2 和最低的 *RMSE*,无论是在训练集还是测试集上都显示出优异的性能和泛化能力。PSO-SVR 模型在大多数监测

断面上也表现良好,其预测性能略逊于 RF 模型,但稳定性和泛化能力都较好,可进一步优化其网络结构或参数来提高预测精度和稳定性。

本研究计算了 2 种模型在训练集和测试集上的 R^2 和 *RMSE*,对鄱阳湖 8 个监测站点机器学习模型预测性能进行评价。如图 7 所示,训练集上整体预测性能为:RF (R^2 = 0.953;*RMSE* = 0.397 mg/L) >PSO-SVR (R^2 = 0.822;*RMSE* = 0.764 mg/L)。模型在预测集上整体预测性能为:RF (R^2 = 0.836;*RMSE* = 0.660 mg/L) > PSO-SVR (R^2 = 0.815;*RMSE* = 0.686 mg/L)。RF 模型 R^2 在训练集和预测集上分布较为集中,表明模型具有较好的稳定性和泛化能力。*RMSE* 在训练集和预测集上也较为集中,但在预测集上略有升高。PSO-SVR 模型的 R^2 和 *RMSE* 在训练集和测试集上的分布较为分散,表明模型在不同断面上的性能变化较大,可能需要针对不同的数据特点进行调整。综合来看,RF 模型在所有监测断面上均表现出最佳的预测能力,具有最高的 R^2 和最低的 *RMSE*,无论是在训练集还是测试集上都显示出优异的性能和泛化能力。PSO-SVR 模型在大多数监测断面上也表现良好,其预测性能略逊于 RF 模型,可能需要进一步优化其网络结构或参数来提高预测精度和稳定性。

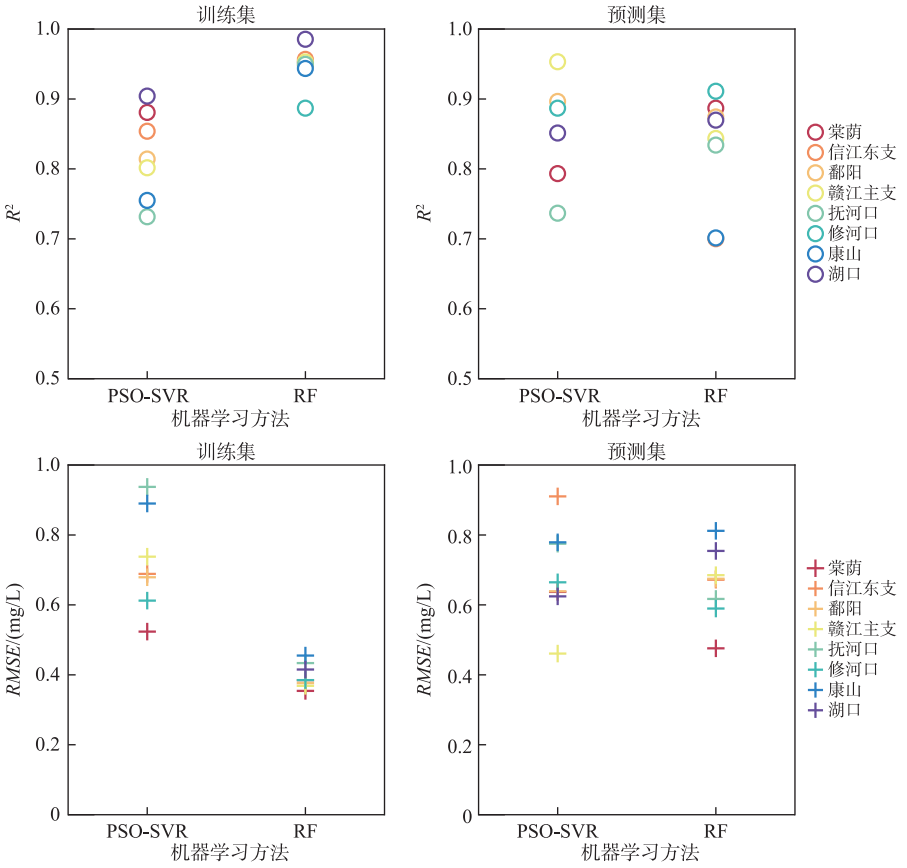

图 7 模型在训练集和预测集上的 R^2 值和 *RMSE* 值分布

Fig.7 Distribution of R^2 and *RMSE* values of the model on the training and prediction sets

3 结论

通过对鄱阳湖长期水质监测数据的系统分析,揭示了 DO 的时空变化特征及其主要影响因素,并通过模型预测分析了未来的水质变化趋势。研究结果对理解鄱阳湖的生态环境变化及其驱动机制具有重要意义,具体结论如下。

1)利用鄱阳湖区及入湖“五河”的实测资料,使用 MK 趋势检验分析了 1988－2023 年鄱阳湖 8 个监测站点的 DO 变化特征。整体上 DO 浓度上升的站点为抚河口、修河口、康山和湖口,其中康山和湖口的 DO 浓度在后期表现出显著上升趋势。对于棠荫、鄱阳、赣江主支、信江东支 4 个站点,由于 UF 和 UB 线没有明显超过±1.96 的阈值,因此没有明显的突变点,呈现出周期性波动变化的特点。这些发现为理解鄱阳湖水质的长期变化趋势提供了重要的科学依据,有助于制定针对性的水质管理和保护措施。

2)基于随机森林重要性指数探究了 DO 与其他水质因子间的响应和关系,在 8 个监测站点中 T 对 DO 的重要性指数均较高,其次是 COD_{Mn},说明 2 者对鄱阳湖 DO 浓度时空分布的影响较大。各个因子的平均 IMI 排名为 T> COD_{Mn}>TN>NH_3-N>TP>pH,其重要性指数值分别为 2.54、0.81、0.65、0.63、0.43 和 0.37。DO 水平显示出明显的季节性变化,夏季较低,冬季升高,这与水温变化、水生植物活动和河流流量有关,这些结果有助于识别和管理影响湖泊水质的关键因素,提高水质管理的针对性和有效性。

3)使用随机森林和改进支持向量机回归对 1988－2023 年月均水质数据进行模型预测对比分析。整体上,RF 和 PSO-SVR 模型在 8 个监测站点的总体平均误差分别为 0.32 和 0.54。基于混淆矩阵的模型性能评价中,RF 和 PSO-SVR 的平均准确率 η 分别为 0.67 和 0.52。在分站点的混淆矩阵结果中,误差低与准确率高之间并不呈现对应的关系,例如 PSO-SVR 在抚河口的预测任务中误差仅为 0.40,而混淆矩阵结果显示抚河口的 PSO-SVR 预测存在着普遍偏高的趋势。基于 R^2 和 *RMSE* 对模型预测性能的评价结果,训练集上整体预测性能为:RF (R^2 = 0.953;*RMSE* = 0.397 mg/L)>PSO-SVR (R^2 = 0.822;*RMSE* = 0.764 mg/L)。模型在预测集上整体预测性能为:RF (R^2 = 0.836;*RMSE* = 0.660 mg/L)>PSO-SVR (R^2 = 0.815;*RMSE* = 0.686 mg/L)。两种模型均表现出优秀的预测性能,其中 RF 的预测能力更好,无论是在训练集还是测试集上都显示出优异的性能和泛化能力。PSO-SVR 模型在大多数监测断面上也表现良好,其预测性能略逊于 RF 模型,但稳定性和泛化能力都较好。综合来看,两种模型在鄱阳湖水质预测中都具有较好的应用前景,为未来水质变化预测提供了可靠的技术支持。

4 参考文献

[1] Ke XZ, Tao YQ, Zhang XX *et al*. Geogenic and anthropogenic impacts on phosphorus enrichment in groundwater around China's largest freshwater lake. *Journal of Hydrology*, 2024, **635**: 131154. DOI: 10.1016/j.jhydrol.2024.131154.

[2] Zhang L, Yuan LJ, Xiang JJ *et al*. Response of the microbial community structure to the environmental factors during the extreme flood season in Poyang Lake, the largest freshwater lake in China. *Frontiers in Microbiology*, 2024, **15**: 1362968. DOI: 10.3389/fmicb.2024.1362968.

[3] Zhu ZZ, Zhang SL, Zhang YR *et al*. Flood risk transfer analysis based on the “Source-Sink” theory and its impact on ecological environment: A case study of the Poyang Lake Basin, China. *Science of the Total Environment*, 2024, **921**: 171064. DOI: 10.1016/j.scitotenv.2024.171064.

[4] Li WX, Jiang ML, Xu LG *et al*. Spatial and temporal characteristics of phytoplankton in Lake Poyang and its response to extreme flood and drying events. *J Lake Sci*, 2024, **36**(4): 1001-1013. DOI: 10.18307/2024.0411. [李文轩, 蒋名亮, 徐力刚等. 鄱阳湖浮游植物时空变化特征及其对极端洪枯事件的响应. 湖泊科学, 2024, **36**(4): 1001-1013.]

[5] Wu Y, Wang C, Wang H *et al*. Analysis of water quality of Le'an River in Poyang Lake Basin based on CCME-WQI method. *Environmental Science*, 2024, **45**(9): 5235-5243. [吴怡, 王成, 王华等. 基于 CCME-WQI 方法的鄱阳湖流域乐安河水质分析. 环境科学,2024, **45**(9): 5235-5243.]

[6] Liu ZJ, Wang XH, Jia SQ *et al*. Eutrophication causes analysis under the influencing of anthropogenic activities in China's largest fresh water lake (Poyang Lake): Evidence from hydrogeochemistry and reverse simulation methods. *Journal of Hydrology*, 2023, **625**: 130020. DOI: 10.1016/j.jhydrol.2023.130020.

[7] Han Q, Meng H, Wang SL *et al*. Impacts of submerged macrophyte communities on ecological health: A comprehensive assessment in the western waters of Yuanmingyuan Park replenished by reclaimed water. *Ecological Engineering*, 2024, **202**: 107220. DOI: 10.1016/j.ecoleng.2024.107220.

[8] Deng GY, Liu Y, Jiang HB *et al*. Disentangling the relative and cumulative impacts of diverse policies on food- and water-related ecosystem services and their trade-offs in ecologically fragile areas. *Journal of Cleaner Production*, 2024, **447**: 141322. DOI: 10.1016/j.jclepro.2024.141322.

[9] Shi XL, Wang LP, Chen A *et al*. Enhancing water quality and ecosystems of reclaimed water-replenished river: A case study of Dongsha River, Beijing, China. *Science of the Total Environment*, 2024, **926**: 172024. DOI: 10.1016/j.scitotenv.2024.172024.

[10] Yin YZ, Xia R, Chen Y *et al*. Non-steady state fluctuations in water levels exacerbate long-term and seasonal degradation of water quality in river-connected lakes. *Water Research*, 2023, **242**: 120247. DOI: 10.1016/j.watres.2023.120247.

[11] Xu L, Hu Q, Jian MF *et al*. Exploring the optical properties and molecular characteristics of dissolved organic matter in a large river-connected lake (Poyang Lake, China) using optical spectroscopy and FT-ICR MS analysis. *Science of the Total Environment*, 2023, **879**: 162999. DOI: 10.1016/j.scitotenv.2023.162999.

[12] Li B, Yang GS, Wan RR *et al*. Impacts of hydrological alteration on ecosystem services changes of a large river-connected lake (Poyang Lake), China. *Journal of Environmental Management*, 2022, **310**: 114750. DOI: 10.1016/j.jenvman.2022.114750.

[13] Xu L, Yuan SY, Tang HW *et al*. Mixing dynamics at the large confluence between the Yangtze River and Poyang Lake. *Water Resources Research*, 2022, **58**(11): e2022WR032195. DOI: 10.1029/2022WR032195.

[14] Chen T, Song CQ, Fan CY *et al*. Remote sensing modeling of environmental influences on lake fish resources by machine learning: A practice in the largest freshwater lake of China. *Frontiers in Environmental Science*, 2022, **10**: 944319. DOI: 10.3389/fenvs.2022.944319.

[15] Wang WY, Yang P, Xia J *et al*. Changes in the water environment and its major driving factors in Poyang Lake from 2016 to 2019, China. *Environmental Science and Pollution Research*, 2023, **30**(2): 3182-3196. DOI: 10.1007/s11356-022-22136-3.

[16] Hou XK, Gao W, Zhang M *et al*. Source apportionment of water pollutants in Poyang Lake Basin in China using absolute principal component score-multiple linear regression model combined with land-use parameters. *Frontiers in Environmental Science*, 2022, **10**: 924350. DOI: 10.3389/fenvs.2022.924350.

[17] WU Y, Wang H, Deng Y *et al*. Evaluation and driving characteristics of water nutrients in Poyang Lake based on TLI method. *Environmental Engineering*, 2024, **42**(5): 10-17. DOI: 10.13205/j.hjgc.202405002. [吴怡, 王华, 邓燕青等. 基于 TLI 法的鄱阳湖水体营养状态评价与驱动特征分析. 环境工程 , 2024, **42**(5): 10-17.]

[18] Mao ZY, Xu LG, Lai XJ *et al*. Assessment on ecosystem health of Lake Poyang based on a comprehensive index method. *J Lake Sci*, 2023, **35**(3): 1022-1032. DOI: 10.18307/2023.0321. [毛智宇, 徐力刚, 赖锡军等. 基于综合指标法的鄱阳湖生态系统健康评价. 湖泊科学, 2023, **35**(3): 1022-1032.]

[19] 徐闯, 余香英, 许泽婷等. 潭江干流溶解氧时空格局及其调控因素研究. 环境科学学报, 2024, **44**(7): 222-230. DOI: 10.13671/j.hjkxxb.2024.0027.

[20] Shen X, Li S, Cai HJ *et al*. The response mechanism of transversal mixing of dissolved oxygen to the evolution of secondary flow at the confluence. *Journal of Hydrology*, 2024, **635**: 131184. DOI: 10.1016/j.jhydrol.2024.131184.

[21] Gao WD, Chen HJ, Ma XY *et al*. Hexavalent chromium removal by green synthesized nano-size iron particles combined with iron sulfides: Effects of dissolved oxygen and phosphate. *Journal of Environmental Chemical Engineering*, 2024, **12**(3): 112673. DOI: 10.1016/j.jece.2024.112673.

[22] Yao JW, Li Y, Lv YJ *et al*. Analysis on water quality evolution trend of drinking water source based on water quality index method and M-K test. *Environmental Science and Management*, 2024, **49**(4): 43-48. [姚嘉伟, 李燕, 吕业佳等. 基于水质指数法和 M-K 检验的饮用水水源地水质演变趋势研究. 环境科学与管理, 2024, **49**(4): 43-48.]

[23] Chi DW, Huang Q, Liu LZ. Dissolved oxygen concentration prediction model based on WT-MIC-GRU—A case study in dish-shaped lakes of Poyang Lake. *Entropy*, 2022, **24**(4): 457. DOI: 10.3390/e24040457.

[24] Xin BD, Lv LH, Wang P *et al*. Spatiotemporal variation characteristics of ozone and identification of key influencing factors based on random forest model: A case study of Chuzhou City. *Environmental Science*, 2024, **45**(9): 5117-5126. [辛泊达, 吕连宏, 王培等. 基于随机森林模型的臭氧浓度时空变化特征及关键影响因子识别: 以滁州市为例. 环境科学, 2024, **45**(9): 5117-5126.]

[25] Yang XT, Kang P, Wang AY *et al*. Prediction of ozone pollution in Sichuan Basin based on random forest model. *Environmental Science*, 2024, **45**(5): 2507-2515. [杨晓彤, 康平, 王安怡等. 基于随机森林模型的四川盆地臭氧污染预测. 环境科学, 2024, **45**(5): 2507-2515.]

[26] Alnahit AO, Mishra AK, Khan AA. Stream water quality prediction using boosted regression tree and random forest models. *Stochastic Environmental Research and Risk Assessment*, 2022, **36**(9): 2661-2680. DOI: 10.1007/s00477-021-02152-4.

[27] Geetha JM. Secure water quality prediction system using machine learning and blockchain technologies. *Journal of Environmental Management*, 2024, **350**: 119357. DOI: 10.1016/j.jenvman.2023.119357.

[28] Sarafaraz J, Ahmadzadeh KF, Mahmoudi KJ *et al*. Predicting river water quality: An imposing engagement between machine learning and the QUAL2Kw models (case study: Aji-Chai, River, Iran). *Results in Engineering*, 2024, **21**: 101921. DOI: 10.1016/j.rineng.2024.101921.

[29] Peng YJ, Chen G, Chao NF *et al*. Detection of extreme hydrological droughts in the Poyang Lake basin during 2021-2022 using GNSS-derived daily terrestrial water storage anomalies. *Science of the Total Environment*, 2024, **919**: 170875. DOI: 10.1016/j.scitotenv.2024.170875.

[30] Mo LM, Wan NN, Zhou B *et al*. Per- and polyfluoroalkyl substances in waterbird feathers around Poyang Lake, China: Compound and species-specific bioaccumulation. *Ecotoxicology and Environmental Safety*, 2024, **273**: 116141. DOI: 10.1016/j.ecoenv.2024.116141.

[31] Chen HX, Zhang ZT, Jin GQ *et al*. Effects of periodic fluctuation of water level on solute transport in seasonal lakes in Poyang floodplain system. *Water Resources Research*, 2023, **59**(12): e2023WR034739. DOI: 10.1029/2023WR034739.

[32] Xu L, Hu Q, Liu ZT *et al*. Hydrological alteration drives chemistry of dissolved organic matter in the largest freshwater lake of China (Poyang Lake). *Water Research*, 2024, **251**: 121154. DOI: 10.1016/j.watres.2024.121154.

[33] Zeng JF, Qiu JF, Wu ZY *et al*. Impact of the Three Gorges Dam on hydrological connectivity and vegetation growth of Poyang Lake floodplain, China. *Journal of Hydrology*, 2024, **631**. DOI: 10.1016/J.JHYDROL.2024.130831.

[34] Xu XL, Zhao J, Wu GD *et al*. Tracing water recharge and transport in the root-zone soil of different vegetation types in the Poyang Lake floodplain wetland (China) using stable isotopes. *Sustainability*, 2024, **16**(5): 1755. DOI: 10.3390/su16051755.

[35] Xia ZM, Liao KT, Guo LP *et al*. Spatiotemporal distribution pattern of precipitation in Ganjiang River Basin based on EOF and MK trend analysis. *Research of Soil and Water Conservation*, 2023, **30**(5): 223-233, 249. [夏志明, 廖凯涛, 郭利平等. 基于 EOF 和 MK 趋势分析的赣江流域降水时空分布格局. 水土保持研究, 2023, **30**(5): 223-233, 249.]

[36] Ma YG, Huang Y. Interannual and seasonal trend analysis of vegetation condition in Xinjiang based on 1982-2013 NDVI data. *Climatic and Environmental Research*, 2018, **23**(1): 26-36. [马勇刚, 黄粤. 基于 1982－2013 年 NDVI 数据的新疆 30 年植被状况季节与年际趋势分析. 气候与环境研究, 2018, **23**(1): 26-36.]

[37] Singh H, Choudhary MP. Trend analysis of rainfall and groundwater level in Jaisalmer District of the Thar Desert, Rajasthan, India. *Water Resources*, 2023, **50**(2): S134-S143. DOI: 10.1134/S0097807822100505.

[38] Tian Y, Wang S, Pei LW *et al*. Electrochemical mechanism of synchronous ammonia and nitrate removal based on multi-objective optimization by coupling random forest with genetic algorithm. *Science of the Total Environment*, 2023, **901**: 166039. DOI: 10.1016/j.scitotenv.2023.166039.

[39] Liu XF, Zhang L, Yang FH *et al*. Determining reclaimed water quality thresholds and farming practices to improve food crop yield: A meta-analysis combined with random forest model. *Science of the Total Environment*, 2023, **862**: 160774. DOI: 10.1016/j.scitotenv.2022.160774.