**Artificial Intelligence and Robotics Research** 人工智能与机器人研究, 2026, 15(6), 1358-1365
Published Online November 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.156123

Hans汉斯

# 基于SG-RF-MTBLSTM模型的湿地水质预测研究

**马滨滨，梁雅晴，何雨薇，叶晓霞***

广东海洋大学数学与计算机学院，广东 湛江

收稿日期：2026年8月31日；录用日期：2026年10月27日；发布日期：2026年11月5日

## 摘　要

**为解决湿地水质预测受潮汐气象等环境、参数耦合及极端天气干扰，本文提出SG-RF-MTBLSTM混合预测模型。对近岸湿地溶解氧(DO)与原位硝酸盐($NO_3$)进行预测，两者受潮汐、气象与参数耦合共同影响：先对连续小时时序作Savitzky-Golay (SG)平滑，经随机森林(RF)得到精简协变量子集，再以共享BiLSTM骨干 + DO/$NO_3$双头，在24 h滑动窗口上同步估计窗末时刻的溶解氧与硝酸盐。基于USGS连续探头站2020~2025年小时数据，站点为USGS-02292900，按时间划分为2020~2022年训练，2023~2024年验证，2025年测试。结果表明，模型测试集DO的$R^2$为0.80，$NO_3$的$R^2$为0.69；以6~11月为飓风季分层后，飓风季DO的MAE由正常期0.59 mg/L升至0.94 mg/L、$R^2$由0.74降至0.65，$NO_3$同步轻度下降，存在季节性分布外泛化困难。**

## 关键词

**湿地水质预测，随机森林，多任务双向长短期记忆网络**

# Research on Wetland Water Quality Prediction Based on the SG-RF-MTBLSTM Model

**Binbin Ma, Yaqing Liang, Yuwei He, Xiaoxia Ye***

College of Mathematics and Computer, Guangdong Ocean University, Zhanjiang Guangdong

Received: August 31, 2026; accepted: October 27, 2026; published: November 5, 2026

*通讯作者。

文章引用: 马滨滨, 梁雅晴, 何雨薇, 叶晓霞. 基于 SG-RF-MTBLSTM 模型的湿地水质预测研究[J]. 人工智能与机器人研究, 2026, 15(6): 1358-1365. DOI: 10.12677/airr.2026.156123

## Abstract

**To address the challenges of coastal wetland water quality prediction under tidal-meteorological forcing, cross-parameter coupling, and extreme-weather disturbances, this paper proposes the SG-RF-MTBLSTM hybrid prediction model. The targets are nearshore dissolved oxygen (DO) and *in-situ* nitrate ($NO_3$), which are jointly driven by tidal, meteorological, and biogeochemical factors. Continuous hourly series are first smoothed with a Savitzky-Golay (SG) filter; a random forest (RF) then yields a compact covariate subset; and a shared bidirectional LSTM (BiLSTM) backbone with dual DO/$NO_3$ heads jointly estimates window-end DO and nitrate over a 24 h sliding window. Experiments use USGS continuous-probe hourly data for 2020~2025 at the primary site USGS-02292900, with a chronological split into training (2020~2022), validation (2023~2024), and testing (2025). On the test set, the model attains an $R^2$ of 0.80 for DO and 0.69 for $NO_3$. After a June-November hurricane-season stratification, hurricane-season DO MAE rises from 0.59 mg/L in the normal period to 0.94 mg/L and $R^2$ falls from 0.74 to 0.65, with a mild concurrent decline for $NO_3$, indicating seasonal out-of-distribution generalization difficulty.**

## Keywords

**Wetland Water Quality Prediction, Random Forest, Multi-Task Bidirectional Long Short-Term Memory Network**

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. 引言与相关工作

水质参数动态监测是湿地生态保护的重要支撑。连续探头监测可提供小时级 DO、水温、电导、pH 及原位硝酸盐等序列，但多参数耦合与季节或者极端天气扰动仍使数据驱动预报具有挑战性。早期预测多依靠 LSTM [1]、随机森林[2]等单一模型，难以兼顾水质参数短期波动情况与长期变化趋势。随着技术发展，无人机遥感与多源传感技术为湿地水质监测提供了新路径[3]，小波去噪 LSTM-Transformer 等模型取得较好效果[4]，但现阶段各方法还是很难充分融合水文、气象以及潮汐等多模态数据，复杂建模能力仍不足，极端事件预判能力有限。

本研究搭建 SG-RF-MTBLSTM 混合模型，通过随机森林自动筛选关键协变量，在 24 h 滑动窗口上以共享 BiLSTM 进行时序建模，依托多任务学习框架[5]，实现溶解氧与硝酸盐的同步估计。基于美国佛罗里达州沿岸 USGS 连续探头 2020~2025 年小时数据，以 USGS-02292900 为主报站进行检验。结果表明：测试集 DO 的 $R^2$ 为 0.80，$NO_3$ 的 $R^2$ 为 0.69；正常期相对稳定，但飓风季 DO 误差显著上升，机制诊断指向驱动关系偏移。本研究为近岸湿地双目标连续监测提供可复现方法，并为极端工况误差分析提供依据。

## 2. 方法

本文提出 SG-RF-MTBLSTM 混合模型，先对原始水质时序进行 Savitzky-Golay 平滑预处理，再经 RF 完成特征筛选，最后输入共享 BiLSTM 实现 DO、$NO_3$ 双目标同步预测。模型架构如图 1 所示。

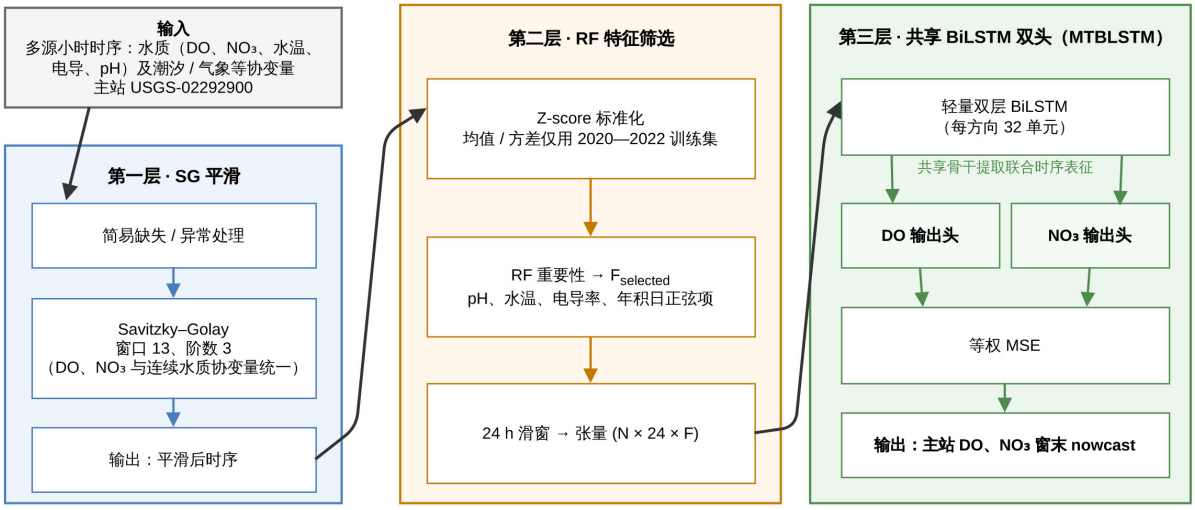



**Figure 1.** Architecture of the SG-RF-MTBLSTM model
**图 1.** SG-RF-MTBLSTM 模型架构图

## 2.1. SG 数据预处理

采用 Savitzky-Golay 滤波对水质时序作局部多项式平滑：窗口长度 13、多项式阶数 3，以抑制高频噪声并保留峰谷形态。

## 2.2. RF 特征工程

### 2.2.1. 候选协变量与 RF 特征筛选

预处理(SG 平滑)之后，在连续小时表上形成候选协变量集合，再由随机森林(RF)按重要性筛选得到精简子集 $F_{selected}$，作为后续共享 BiLSTM 的输入(见图 2)。

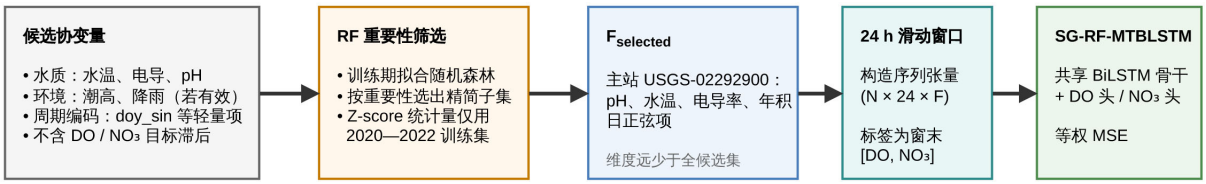



**Figure 2.** Architecture of the feature engineering framework
**图 2.** 特征工程体系架构图

(1) 候选协变量

候选变量主要包括水质连续观测参数，即水温、电导率、pH、环境对齐量，即潮高、降雨以及轻量周期编码，即年积日正弦项；明确排除溶解氧(DO)与硝酸盐($NO_3$)的滞后项，以避免目标泄漏式自回归。

(2) 标准化

对数值型候选特征，仅用训练集(2020~2022)估计均值与标准差，做 $Z$-score：

$$Z = \frac{X - \mu_{\text{train}}}{\sigma_{\text{train}}}, \tag{1}$$

(3) RF 筛选

在训练期(2020~2022)，对候选协变量 $\boldsymbol{z} \in \mathbb{R}^F$ 分别以 DO、$NO_3$ 为因变量拟合随机森林回归。目标

$k \in \{\mathrm{DO}, \mathrm{NO}_3\}$ 的 RF 预测为 $M$ 棵树的平均( $M = 200$ )：

$$\hat{y}_k^{\mathrm{RF}} = \frac{1}{M} \sum_{n=1}^{M} f_n^{(k)}(z), \tag{2}$$

其中 $f_n^{(k)}$ 表示针对目标 $k$ 训练的随机森林中的第 $n$ 棵决策树。

由不纯度下降得到特征 $j$ 的重要性 $\mathrm{Imp}_j^{(k)}$，并归一化：

$$\widetilde{\mathrm{Imp}}_j^{(k)} = \frac{\mathrm{Imp}_j^{(k)}}{\sum_{j'=1}^{F} \mathrm{Imp}_{j'}^{(k)}} \tag{3}$$

按重要性降序取前 $K$ 个特征， $K = \rho F$ (本文 $\rho = 0.40$ )：

$$\mathcal{F}^{(k)} = \mathrm{Top}_K\left(\widetilde{\mathrm{Imp}}^{(k)}\right), \tag{4}$$

双目标最终入模子集取并集：

$$\mathcal{F}_{\mathrm{selected}} = \mathcal{F}^{(\mathrm{DO})} \cup \mathcal{F}^{(\mathrm{NO}_3)}, \tag{5}$$

主站 USGS-02292900 终跑所用 $\mathcal{F}_{\mathrm{selected}}$ 为：pH、水温、电导率、年积日正弦项。

#### 2.2.2. 时间序列构建

以每个监测站点为单位，将连续时间线切成固定长度滑动窗口，默认 $W = 24$ h 。每个窗口仅包含该时段内筛选后的协变量 $\mathcal{F}_{\mathrm{selected}}$ (不含 RF 辅助预测)，构成三维张量：

$$S_t = \left[X_{t-W+1}, X_{t-W+2}, \cdots, X_t\right] \in \mathbb{R}^{W \times F}, F = \left|\mathcal{F}_{\mathrm{selected}}\right|, \tag{6}$$

序列标签为窗口末端两目标真实值：

$$y_t = \left[\mathrm{DO}_t, \mathrm{NO}_{3,t}\right], \tag{7}$$

### 2.3. 多任务双向 LSTM 预测网络

在得到精简协变量子集 $\mathcal{F}_{\mathrm{selected}}$ 并完成 24 h 滑窗构造后，本文将序列张量 $S_t \in \mathbb{R}^{W \times F}$ ， $W = 24$ ， $F = \left|\mathcal{F}_{\mathrm{selected}}\right|$ 输入多任务双向长短期记忆网络(MTBLSTM)，同步估计窗末时刻的溶解氧与硝酸盐。记标签为 $y_t = \left[\mathrm{DO}_t, \mathrm{NO}_{3,t}\right]$ ，即对同一时刻 $t$ 的双目标同步估计。

(1) 共享骨干与双头输出网络核心为两层双向 LSTM：第一层返回完整时间步隐状态，第二层输出窗口末端的双向拼接表征；每层每方向隐单元数为 32。双向结构分别沿时间正向与反向提取依赖，从而同时刻画短期波动与日周期内的前后文信息。两层 BiLSTM 之后，经一层 ReLU 全连接(128 单元)与轻度 Dropout (丢弃率 0.20)，再分支为两个线性回归头，分别输出标准化空间中的 $\widehat{\mathrm{DO}}_t$ 与 $\widehat{\mathrm{NO}}_{3,t}$ 。两头共享同一骨干参数，使 DO 与 $\mathrm{NO}_3$ 在隐空间中共享时序表示。

(2) 目标标准化与等权损失因 DO 与 $\mathrm{NO}_3$ 量纲与方差差异较大，训练阶段对两目标分别用训练集拟合 Z-score 标准化，在标准化空间计算损失；评估时再反变换回物理浓度(mg/L)。总损失取等权均方误差：

$$\mathcal{L} = \omega_{\mathrm{DO}} \mathrm{MSE}(\mathrm{DO}) + \omega_{\mathrm{NO}_3} \mathrm{MSE}(\mathrm{NO}_3), \omega_{\mathrm{DO}} = \omega_{\mathrm{NO}_3} = 1, \tag{8}$$

## 3. 实验

水质数据来自美国地质调查局(USGS)国家水信息系统(NWIS)连续探头小时观测，主站为 USGS-02292900，参数含溶解氧、水温、电导率、pH 及原位硝酸盐 + 亚硝酸盐(参数代码 99133)，时段

2020~2025，可通过 NWIS 按站号与参数代码检索下载。潮汐数据来自美国国家海洋和大气管理局(NOAA)潮汐与海流门户(Tides and Currents, https://www.tidesandcurrents.noaa.gov)的水位产品，对齐至小时潮高。气象要素来自佛罗里达自动化气象网络(FAWN，Florida Automated Weather Network，佛罗里达大学 IFAS，https://fawn.ifas.ufl.edu)区域站点日汇总归档(含降雨等)，按时间戳对齐至小时；划分：训练集为 2020~2022 年，验证集为 2023~2024 年，测试集为 2025 年。优化器为 Adam (初始学习率 $1 \times 10^{-3}$)，最大训练轮数为 60，并配合早停(patience = 8)、验证损失平台期学习率衰减(衰减因子 0.50)及按验证损失保存最优权重；LSTM 层 Dropout 取 0.15，流程见图 3。基线为 LSTM、RF 与 ARIMA。评价指标包括 MSE、MAE、RMSE、$R^2$ 与 sMAPE，均在反标准化后的物理浓度上计算。实验在 AutoDL 云主机完成，环境为 Ubuntu 20.04、Python 3.8、TensorFlow 2.9.0 与 CUDA 11.2，硬件为 1 × NVIDIA GeForce RTX 3080 Ti (12 GB)。

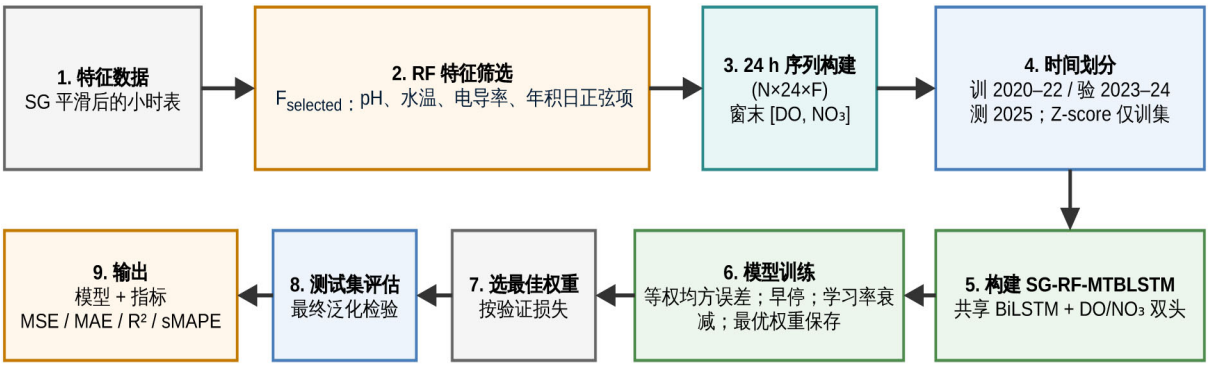



**Figure 3.** Model training and optimization workflow
**图 3.** 模型训练与优化流程框图

## 4. 实验分析

模型评估指标包括均方误差(MSE)、平均绝对误差(MAE)、均方根误差(RMSE)、决定系数($R^2$)及对称平均绝对百分比误差(sMAPE)。主表报告测试集(2025)整体表现，并与无共享基线对照；分层评估以 6~11 月为飓风季、其余为正常期，在同一套主模型测试预测上统计误差变化。

### 4.1. 整体预测性能分析

由表 1 可见，SG-RF-MTBLSTM 在双目标预测任务上整体优于基线模型。针对溶解氧(DO)，本文模型决定系数 $R^2$ 为 0.80，高于 LSTM (0.74)与随机森林 RF (0.71)；同时平均绝对误差(MAE)、均方根误差(RMSE)与对称平均绝对百分比误差(sMAPE)均为最低。自回归积分滑动平均模型(ARIMA)的 $R^2$ 为负值，难以刻画多因子驱动的小时尺度时序变化。对于硝酸盐($NO_3$)，模型优势更为突出，本文模型 $R^2 = 0.69$，显著高于 LSTM (0.32)与 RF (0.53)。对于数值分析易得，$NO_3$ 明显比 DO 更难估，因为 DO 与水温、pH、电导及昼夜代谢过程耦合较紧，主站入模特征对其解释力更强，$NO_3$ 还受硝化/反硝化、潮汐混合与营养脉冲等生物地球化学过程影响，与现有精简协变量的映射更弱、更非线性。而且 $NO_3$ 浓度量级远低于 DO，同等绝对误差会放大相对指标，如 sMAPE。此外，本站对齐后小时降雨有效变幅不足，难以直接刻画冲刷–稀释等关键外强迫，进一步抬高了 $NO_3$ 的可预报难度。而共享双向长短期记忆网络(BiLSTM)骨干的主要收益，正在于提升并稳定这一更难的一侧，而不仅是抬高 DO。但全年较高的 $R^2$ 并不代表模型在各季节均具备同等稳健性，飓风季分层分析详见 4.2 节。

**Table 1.** Overall model performance metrics
**表 1.** 模型整体性能指标

| (a) 溶解氧 DO | | | | | |
|---|---|---|---|---|---|
| 模型 | MSE | MAE | RMSE | $R^2$ | sMAPE (%) |
| SG-RF-MTBLSTM | 0.98 | 0.78 | 0.99 | 0.80 | 6.53 |
| LSTM | 1.35 | 0.93 | 1.16 | 0.74 | 9.21 |
| RF | 1.49 | 0.86 | 1.22 | 0.71 | 6.58 |
| ARIMA | 5.10 | 1.83 | 2.26 | −0.04 | 14.34 |
| (b) 硝酸盐 $NO_3$ | | | | | |
| 模型 | MSE | MAE | RMSE | $R^2$ | sMAPE (%) |
| SG-RF-MTBLSTM | 0.00567 | 0.06 | 0.08 | 0.69 | 15.50 |
| LSTM | 0.00949 | 0.08 | 0.10 | 0.32 | 24.66 |
| RF | 0.00663 | 0.06 | 0.08 | 0.53 | 16.61 |
| ARIMA | 0.02820 | 0.15 | 0.17 | −0.54 | 29.20 |

## 4.2. 飓风季适应性与误差机制

本文对极端天气相关样本采用评测分层。定义上，将测试集按大西洋飓风季常用约定划分为飓风季(6~11 月)与正常期(其余月份)，该划分用于诊断飓风季和暖季误差。主站测试集正常期与飓风季样本量分别为 3858 与 4392。飓风季相对正常期，如表 2，DO 的 MAE 由 0.59 升至 0.94 mg/L (约+59%)，$R^2$ 由 0.74 降至 0.65，而 $NO_3$ 的 MAE 由 0.05 升至 0.07，$R^2$ 由 0.69 降至 0.61，呈同步轻度变差。

**Table 2.** Normal period vs. hurricane season (June-November)
**表 2.** 正常期 vs. 飓风季(6~11 月)

| 时段 | DO MAE | DO $R^2$ | $NO_3$ MAE | $NO_3$ $R^2$ |
|---|---|---|---|---|
| 正常期 | 0.59 | 0.75 | 0.05 | 0.69 |
| 飓风季 | 0.94 | 0.65 | 0.07 | 0.61 |
| 差(飓 − 正) | +0.35 | −0.10 | +0.02 | −0.08 |

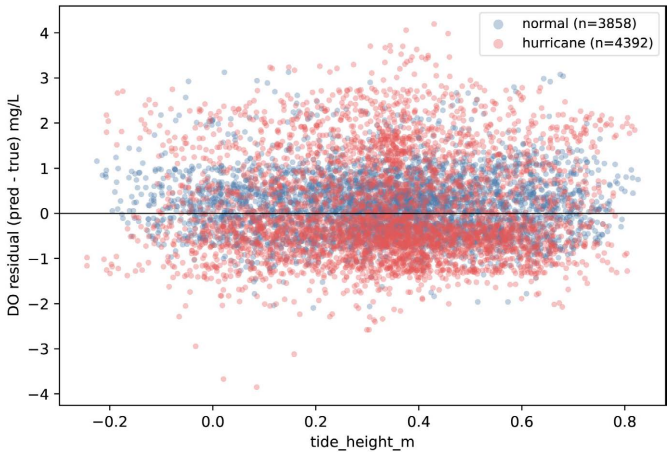



**Figure 4.** DO residual vs. tide level
**图 4.** DO 残差 vs. 潮位

而溶解氧在飓风季预测退化的研究中，因为飓风季样本数 $n$ = 4392 并不少于正常期 $n$ = 3858，因此不宜将误差上升简单归因于极端样本不足。残差 - 潮位散点进一步显示(见图 4)，飓风季残差离散由约±1 mg/L 扩大至约±3~4 mg/L，说明模型在误差系统性增大，而非个别离群点所致。

对 DO (图 5)，其中，图中 ph 为 pH，water_temp_c 为水温(℃)，doy_sin 与 doy_cos 分别为年积日正弦、余弦，hour_sin 与 hour_cos 分别为小时正弦、余弦，tide_height_m 为潮高(m)，rain_mm 为降雨量(mm)，conductivity_us_cm 为电导率(μS/cm)；permutation importance 即置换特征重要性，通过随机打乱单个特征观察模型预测误差增幅衡量特征重要性，mean |MAE| increase 代表平均绝对误差(MAE)增加量的平均值。

进一步用全协变量随机森林置换重要性，对比全部测试与飓风季子集(图 5、图 6)，两期均以 pH 为首要驱动，水温在全部测试中仍有贡献，但在飓风季子集上重要性近乎消失。季节项年积日余弦在飓风季可出现负向重要性，提示季节编码在该子集不稳定。这表明训练期可依赖的水温-DO 及部分季节统计关系在飓风季发生偏移，属协变量——目标条件分布改变，即季节性分布外泛化，共享 BiLSTM 表征外推不足。对 $NO_3$ (图 6)，变化模式与 DO 不同——全部测试上 pH 与年积日正弦更突出，飓风季上二者重要性明显下降，而水温重要性上升；雨、潮、小时周期项两期仍接近零。说明飓风季冲击并非笼统的全局扰动，而是目标特异的驱动关系漂移：DO 侧表现为水温贡献崩塌，$NO_3$ 侧则更多转向依赖水温，同时削弱 pH/季节正弦项。

综上，溶解氧飓风期显著退化主要机制为驱动关系季节性漂移导致的分布外误差放大，而非单纯数据量或单一气象通道失效。

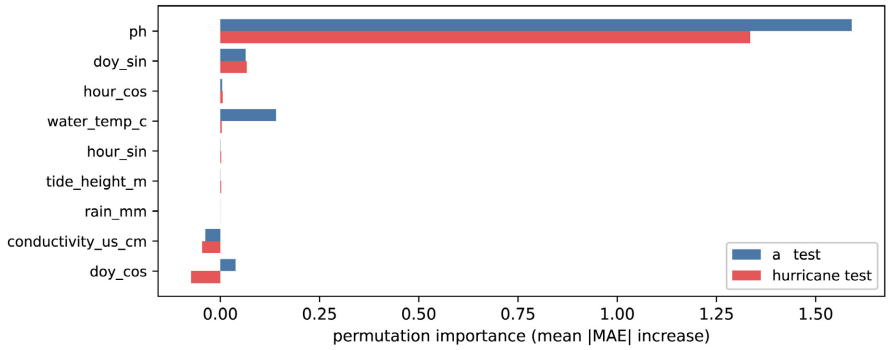



**Figure 5.** Feature importance for DO target
**图 5.** DO 目标的特征重要性

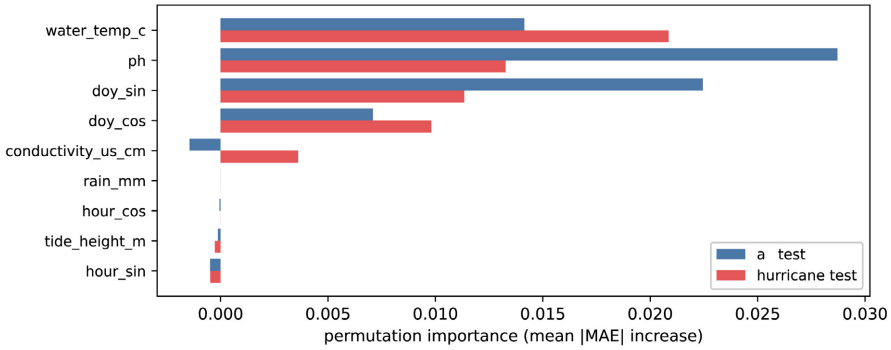



**Figure 6.** Feature importance for $NO_3$
**图 6.** $NO_3$ 目标的特征重要性

## 5. 结论

本文提出面向连续小时监测的 SG-RF-MTBLSTM，以共享 BiLSTM 骨干同步估计窗末 DO 与 $NO_3$；基线为 LSTM、RF 与 ARIMA。主站 USGS-02292900 上，测试 DO 的 $R^2$ 为 0.80，$NO_3$ 的 $R^2$ 为 0.69。飓风季分层显示 DO 显著退化，RF 置换重要性诊断指向水温等驱动关系偏移。未来工作针对上述不足，将采用 iTransformer 在变量维上建模协变量耦合来服务 DO/$NO_3$ 双目标[6]。另在同流域或空间邻近的 USGS 站之间按地理或河网距离构图，并用图神经网络传递站点表征，以缓解单站训练对异质站 $NO_3$ 外推不足。

## 基金项目

湛江市科技攻关计划项目：基于多源数据融合和轻量型 CU-Net 的红树林生态监测系统(NO: 2025B01065)。

## 参考文献

[1] Liang, Z., Zou, R., Chen, X., *et al.* (2020) Simulate the Forecast Capacity of a Complicated Water Quality Model Using the Long Short-Term Memory Approach. *Journal of Hydrology*, **581**, Article 124432. https://doi.org/10.1016/j.jhydrol.2019.124432

[2] Xu, J., Xu, Z., Kuang, J., *et al.* (2021) An Alternative to Laboratory Testing: Random Forest-Based Water Quality Prediction Framework for Inland and Nearshore Water Bodies. *Water*, **13**, 3262. https://doi.org/10.3390/w13223262

[3] 董玉萍, 于泉洲, 宋军, 等. 融合无人机与卫星遥感的南四湖菹草生物量制图及时空格局[J]. 湿地科学与管理, 2025, 21(5): 17-25.

[4] 许轲桐, 曾鸿滨, 陈丽萍, 等. 基于小波去噪的 LSTM-Transformer 模型在河流断面水质预测分析中的应用研究[J]. 环境工程, 2026, 44(3): 125-135.

[5] Ruder, S. (2017) An Overview of Multi-Task Learning in Deep Neural Networks. arXiv preprint arXiv:1706.05098.

[6] Liu, Y., Hu, T., Zhang, H., *et al.* (2024) iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. *International Conference on Learning Representations*, Vienna, 7-11 May 2024, 11116-11140.