

基于时间累积效应与随机森林的巢湖蓝藻水华暴发面积短期预报*

唐晓先¹, 张凯莉^{2,3**}, 段洪涛^{2,3,4}, 邱银国³, 焦亚沁^{2,3}, 罗菊花^{3,4}

(1: 安徽省巢湖管理局湖泊生态环境研究院, 合肥 230071)

(2: 西北大学城市与环境学院, 西安 710127)

(3: 中国科学院南京地理与湖泊研究所湖泊与流域水安全全国重点实验室, 南京 211135)

(4: 中国科学院大学南京学院, 南京 211135)

摘要: 蓝藻水华已成为威胁湖泊生态安全和饮用水安全的全球性环境问题, 及时、准确地预报蓝藻暴发有助于提前采取应对措施, 减轻灾害风险。本研究针对传统机理模型参数众多、运算复杂等应用局限, 以巢湖为研究对象, 构建了融合监测数据与遥感信息的机器学习预报框架。通过整合多站点气象、水质监测数据以及卫星遥感数据, 分析了气象和水质变量对蓝藻水华的时间累积效应。在此基础上, 基于随机森林模型, 分别构建了考虑变量时间累积效应的预报模型(累积变量模型)和仅使用单日观测值的预报模型(单日变量模型), 实现了藻华暴发面积 1~7 天(d)预报。最后, 引入基于博弈论的可解释性(SHapley Additive exPlanations, SHAP)算法, 揭示了主要影响因子的贡献度及其非线性阈值规律。结果表明:(1)气象因子(气温、湿度、降雨、气压)累积效应周期约为 15~30 d, 长于水质因子(氮、磷、溶解氧)的累积效应周期(1~10 d);(2)累积变量模型的预报精度(决定系数 $R^2 = 0.66\sim 0.75$)优于单日变量模型($R^2 = 0.56\sim 0.63$), 其中 1 d 预报效果最优($R^2 = 0.75$, 均方根误差(RMSE) = 49.37 km²);(3)关键阈值条件包括: 平均气温约 > 23 °C、最大风速约 < 4 m/s、降雨量约 > 200 mm、氮磷比约 < 15、pH 约 > 8.5、溶解氧约 < 8.9 mg/L。本研究提出的预报方法仅依赖常规监测数据即可实现短期蓝藻水华预报, 这为富营养化湖泊管理提供了可推广的技术路径与决策支持。

关键词: 蓝藻水华; 时间累积效应; 随机森林; SHAP 可解释性; 短期预报

Integrating temporal cumulative effects into the random forest model for short-term forecasting of cyanobacterial bloom area in Lake Chaohu*

Qang Xiaoxian¹, Zhang Kaili^{2,3**}, Duan Hongtao^{2,3,4}, Qiu Yin'guo³, Jiao Yaqin^{2,3}, Luo Juhua^{3,4}

(1: Institute of Lake Ecology and Environment, Anhui Provincial Lake Chaohu Administration, Hefei 230071, P.R. China)

(2: College of Urban and Environmental Sciences, Northwest University, Xi'an 710127, P.R. China)

(3: State Key Laboratory of Lake and Watershed Science for Water Security, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing 211135, P.R. China)

(4: University of Chinese Academy of Sciences, Nanjing (UCASNJ), Nanjing 211135, P.R. China)

Abstract: Cyanobacterial blooms pose a global environmental challenge, threatening lake ecosystem security and drinking water safety, making timely prediction of their outbreaks critical for implementing preventive measures and reducing associated risks. To address the limitations of conventional mechanism-driven models, which often require numerous parameters and involve high computational complexity, this study developed a machine learning framework integrating multi-source in-situ monitoring and

*2025-04-07 收稿; 2025-10-29 收修改稿。

安徽省巢湖管理局国开行信息化建设“数字巢湖”项目资助。

** 通 xin1 作者; E-mail: 202310268@stumail.nwu.edu.cn

remote sensing data for Lake Chaohu. By combining multi-site meteorological and water quality measurements with satellite-derived time series, we investigated the temporal cumulative effects of meteorological and water quality variables on bloom dynamics. Based on the Random Forest (RF) algorithm, two forecasting models were constructed: one incorporating temporally cumulative variables and the other using only same-day observations, both aimed at predicting bloom coverage area 1–7 days (d) in advance. Furthermore, SHapley Additive exPlanations (SHAP) analysis was employed to interpret the model's decision-making process, revealing feature contributions and nonlinear threshold effects. The results indicate that: (1) meteorological variables (air temperature, humidity, precipitation, and air pressure) exhibited longer cumulative effect durations (15–30 days) than water quality variables (nitrogen, phosphorus, and dissolved oxygen (1–10 days)); (2) cumulative-variable models achieved higher prediction accuracy ($R^2 = 0.7\text{--}0.8$) compared to single-day variable models ($R^2 = 0.4\text{--}0.6$), with the best performance observed for 1-day ahead forecasts ($R^2 = 0.79$, RMSE = 35.36 km²); (3) critical thresholds were identified for average temperature (> 23°C), maximum wind speed (< 4 m/s), precipitation (> 200 mm), nitrogen-phosphorus ratio (< 15), pH (> 8.5), and dissolved oxygen (< 8.9 mg/L). The proposed approach enables high-accuracy short-term forecasting of cyanobacterial blooms using multi-station monitoring data, offering a transferable decision-support framework for the management of eutrophic lakes.

Keywords: Cyanobacterial blooms; temporal cumulative effects; random forest; SHAP interpretability; Short-term forecasting

湖泊作为重要的水体资源,在调节气候、水源供应、生物多样性保护以及促进经济发展方面具有重要作用^[1]。近年来,在全球气候变化与人类活动的双重影响下,湖泊富营养化进程显著加速,导致蓝藻水华频繁发生^[2,3]。蓝藻水华不仅引发水体变色和异味等问题,还会通过藻类过度生长和分解大量消耗溶解氧,对水生生态系统构成严重威胁^[4-6]。此外,蓝藻水华产生的毒素不仅污染水源,还危及动物和人类健康。例如,1996年巴西透析水污染事件导致76人死亡^[7];2007年太湖蓝藻暴发引发无锡市供水危机^[8];2013年美国伊利湖蓝藻水华造成2000名居民断水^[9,10];以及2020年非洲330头大象因藻毒素中毒死亡^[11]。甚至有研究认为,在未来几十年内,蓝藻水华的发生规模和频率可能进一步加剧^[12,13]。然而,由于蓝藻水华具有高度的动态性和复杂性,开发高效的预测模型仍面临重大挑战。因此,解析蓝藻水华发生的影响因子,建立精准的预测体系具有重要的现实意义。

在蓝藻水华预测的研究中,主要采用传统统计方法、过程机理模型以及机器学习模型等不同的技术手段。其中,统计方法如线性回归^[14]、时间序列回归^[15],计算简便,并曾有研究利用多元线性回归来预测伊利湖每年最大藻华程度^[16]。然而,受限于数据平稳性要求和线性假设,该类方法对蓝藻水华中常见的非线性动态特征捕捉能力有限。机理模型,如EFDC、PCLake、MIKE系列等^[17,18],基于物理、化学和生物学原理,能够模拟蓝藻水华形成机制与时空动态。然而,该类模型也存在局限性。一方面,其依赖大量观测数据以确定模型参数和边界条件。例如,在预测尚普兰湖(Lake Champlain)米西斯夸湾(Missisquoi Bay)的藻华时,数据收集和准备工作繁琐,导致模型在极端天气条件下的响应能力受限^[19];另一方面,机理模型缺乏灵活性和通用性。在面对不同湖泊生态系统和环境条件时,常需大量参数调整和重新校准,增加了模型在多变水体环境中的预测复杂性。相比之下,随着人工智能技术的发展,机器学习模型如随机森林(RF)、梯度提升决策树(GBDT)、极端梯度提升(XGBoost)、人工神经网络(ANN)和长短期记忆(LSTM)网络等,因其高效且强大的非线性建模能力,在蓝藻水华预测中展现出显著优势^[20-22]。其中,RF模型通过集成学习和随机抽样策略,在捕捉复杂数据模式的同时,提升了模型的泛化能力,能够适应不同湖泊特点和数据特征,实现个性化训练。与此同时,RF具有较高的计算效率,已成为该领域的常用模型^[23,24]。

尽管RF等机器学习模型在蓝藻水华预测中展现出良好的应用潜力,但性能表现不仅依赖于算法结构,还与输入数据的质量密切相关。因此,选择与构建有效的输入变量是影响预测效果的一个关键环节。现有研究虽已识别出营养盐、温度、光照、风速等关键环境因素^[25,26],但大多数模型仍主要依赖于单日观测数据作为输入。实际上,蓝藻细胞的生长、繁殖直至蓝藻水华形成,实质上是环境变量在一段时间内累积的体现。尽管已有研究关注气象因素的累积效应,指出温度在最大藻华前25~30d、降雨在前10d的影响最为显著^[27]。但水质作为影响蓝藻水华的直接因素,特别是氮和磷等重要营养盐,却因逐日数据获取困难,在累积效应研究中很少被考虑。此外,变量之间的交互作用及变量与蓝藻水华之间的依赖关系也尚未在预

测模型中充分体现。再此背景下，基于博弈论的 SHAP 可解释性方法为解析“黑箱”模型的决策机制提供了有效途径，有助于理解环境变量与蓝藻水华之间的复杂响应关系^[28, 29]。因此，综合考虑各类变量、设计合理的变量累积和组合策略，并引入可解释性分析方法，有望提升蓝藻水华预测的准确性和可靠性。

巢湖是中国第五大淡水湖，因快速城市化和农业集约化，已成为典型的富营养化湖泊，蓝藻水华现象严重^[30, 31]。本研究以巢湖为研究区，整合多站点逐日气象与水质监测数据作为环境变量，将 MODIS 卫星衍生的蓝藻水华面积作为预报目标，构建了基于 RF 模型的蓝藻水华暴发面积短期预报模型。具体目标为：

(1) 分析蓝藻水华与环境变量之间的时间累积效应，据此构建最佳变量组合作为模型输入；(2) 提前 1 d、3 d 以及 7 d 预报蓝藻水华暴发面积，并确定最佳预报时效。(3) 评估变量的相对重要性，阐明关键变量与蓝藻水华之间的依赖关系。本研究通过精细化环境变量时间累积效应并结合机器学习技术，构建了一种仅依赖常规监测数据的短期蓝藻水华预报方法，为富营养化湖泊的水质预报与管理提供了新的视角。

1 研究区与数据集

1.1 巢湖概况

巢湖 (31°25'28" ~ 31°43'28"N, 117°16'54" ~ 117°51'46"E, 图 1)，位于安徽省合肥市，属长江下游左岸水系，为中国第五大淡水湖。湖泊总面积约 770 km²，平均水深 2.89 m，流域面积约 1.35 万平方公里。流域内水系发达，入湖河流主要包括南淝河、十五里河、塘西河、派河、杭埠河、白石天河、柘皋河等，并主要通过裕溪河进入长江。流域内农业和城市面源污染输入，尤其是氮、磷等营养盐的持续积累，是导致湖泊富营养化的主要因素。富营养化状况进一步引发频繁的蓝藻水华，对区域水环境管理构成重大挑战。

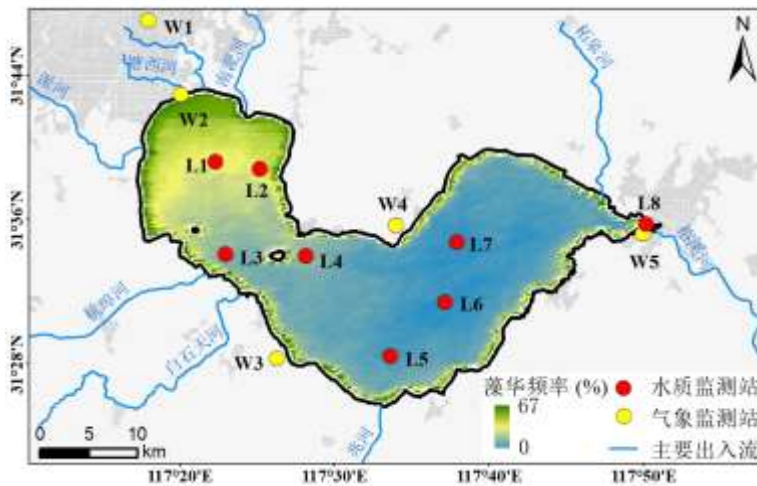


图 1 巢湖及周边气象水质监测站点分布

Fig.1 Distribution of meteorological and water quality monitoring stations in and around Lake Chaohu

1.2 数据收集与处理

1.2.1 蓝藻水华数据集 高频卫星数据对蓝藻水华动态变化监测至关重要。MODIS 卫星提供每日遥感影像，是监测蓝藻水华的关键工具。Ma 等^[32]利用 Terra/MODIS 卫星数据有效提取了 2000 年以来的全球 161 个大型湖泊的蓝藻水华，其中包括巢湖(Ma et al. 2023)。基于该产品，本研究筛选了 2010—2024 年间的 933 幅无云的巢湖蓝藻水华数据 (图 2)，并根据蓝藻水华像元数量计算面积，作为模型的目标变量。并且在模型训练前，为确保数据符合正态分布，对其进行了 log₁₀ 变换。

1.2.2 环境变量集 环境变量包括 2010—2024 年期间的 9 个气象变量：逐日平均气温、最高气温、最低气温、平均风速、最大风速、日照时数、相对湿度、降水量和气压，以及 5 个水质变量：总氮、总磷、氮磷比、pH 和溶解氧。针对部分时段水质数据缺失的情况，采用时间线性插补方法生成逐日连续序列。此外，

考虑到蓝藻水华不仅受总氮和总磷绝对浓度的影响,还可能受其相对比例调控^[33],因此,本研究将氮磷比纳入变量集。数据站点位置见图 1。气象数据中,合肥气象站(W1, #58321)和巢湖气象站(W5, #58326)的数据由中国气象局(<https://data.cma.cn/>)提供,其余气象站(W2~W4)及水质监测站(L1~L8)数据来源于安徽省巢湖管理局数字巢湖平台(<https://chglj.hefei.gov.cn/>)和中国环境监测总站(<https://www.cnemc.cn/>)。此外,需要指出的是,为综合反映巢湖的气象与水质整体状况,本研究对气象监测站(W1~W5)及水质监测站(L1~L8)的逐日数据分别计算算术平均值,作为区域综合代表值。

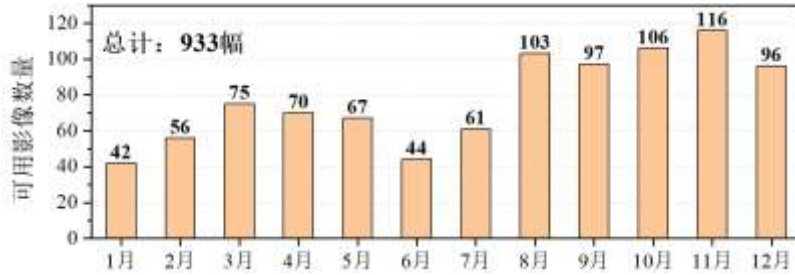


图 2 2010—2024 年无云 Terra/MODIS 卫星影像的月度数量分布

Fig.2 Monthly distribution of cloud-free Terra/MODIS satellite imagery from 2010 to 2024

1.2.3 时间累积变量构建 为权衡预报的时效性和准确性,提前 1 至 7 d 的预报常被认为是合理的时间范围^[22,34,35]。为简化应用程序,本研究选择在目标日期(t)的前 1 d (t-1)、前 3 d (t-3)和前 7 d (t-7)三个日期分别构建预报模型。对于每个预报时效(即 1/3/7 d),模型的输入变量均基于该预报时效之前的历史观测数据进行构建(表 1)。

具体而言,本研究为各环境变量构建了历史累积变量作为模型的输入特征。考虑到在适宜的营养和气象条件下,蓝藻可以持续快速生长一个月^[27],本研究设定 30 d 作为累积变量的最长回溯时间长度。各累积时长所对应的回溯时间区间见表 1 所示。具体构建方法如下:分别从第 t-1 d、第 t-3 d 以及第 t-7 d 起,回溯历史数据,并选择 7 组固定的时间长度(1、5、10、15、20、25 和 30 d)进行累积计算,计算各变量的累积平均值(由于降雨事件通常是间歇性的,且大部分观测日的降雨量为零,降雨量计算为累积总量)。

表 1 1/3/7d 预报时效所对应的累积变量与单日变量的回溯起始-截止时间范围定义。

Tab.1 Definition of the start and end time ranges for the cumulative variables and single-day variables corresponding to the 1/3/7-day forecast.

变量分组	1d 预报 回溯区间	3d 预报 回溯区间	7d 预报 回溯区间	目标预报日期	累积时间长度 (d)
单日变量	t-1	t-3	t-7	t	1
	t-1	t-3	t-7		1
	t-5~t-1	t-7~t-3	t-11~t-7		5
	t-10~t-1	t-12~t-3	t-16~t-7		10
累积变量	t-15~t-1	t-17~t-3	t-21~t-7	t	15
	t-20~t-1	t-22~t-3	t-26~t-7		20
	t-25~t-1	t-27~t-3	t-31~t-7		25
	t-30~t-1	t-32~t-3	t-36~t-7		30

1.2.4 最佳累积变量选择 为确定每个变量的最佳累积时间,本研究采用 Spearman 相关性分析,计算蓝藻水华面积与各组累积变量子集的相关性。在每个变量中,选择具有最大绝对相关系数且通过显著性检验(P < 0.05)的累积变量作为其最佳累积表征。为了降低变量间的冗余和相关性,对温度系列(平均气温、最

高气温、最低气温)、风速系列(平均风速、最大风速)以及氮磷系列(总氮、总磷、氮磷比),分别从每一系列中选取与蓝藻水华面积相关性最强的单一变量参与后续建模。

2 方法

2.1 基于变量穷举组合的 RF 模型

2.1.1 RF 模型原理 RF 是一种集成学习方法,其核心思想是“集体决策”,即通过构建多棵决策树并汇总各树的预测结果来完成学习任务^[36]。为提升模型的泛化能力并降低过拟合风险,RF 引入了两种随机性: **Bootstrap** 抽样和随机特征子集选择,这些特性使得 RF 模型在处理复杂非线性关系和高维特征时表现出色,因而在环境建模与预测等领域广泛应用^[37]。除此之外,尽管基于多层感知器(MLP)、长短期记忆网络(LSTM)、门控循环单元(GRU)及混合架构的先进深度学习模型能有效捕捉蓝藻水华的时间依赖性和复杂动态变化,但这些模型通常面临较高的计算复杂度,并且对超参数设置较为敏感^[38,39]。相比之下,RF 模型的调参过程相对简单,且在模型简洁性、预测准确性及稳健性之间能够提供良好的平衡,这对于环境建模而言具有重要优势。因此,经综合考虑,本研究采用 RF 模型进行蓝藻水华面积的短期预报。

RF 模型的核心原理为:(1) **Bootstrap** 抽样:从原始训练集中有放回地抽取 M 个子样本集,每个子集用于训练一棵决策树。未被抽中的样本构成袋外(Out-of-Bag, OOB)数据集,用于模型内部验证。(2) **随机特征选择**:在每棵树的节点分裂过程中,随机选择一个特征子集,并基于基尼指数或信息增益等准则选择最优分裂点。(3) **决策树生长**:每棵决策树生长至最大深度或不纯度不再显著降低为止,通常不进行后剪枝。(4) **集成结果**:对于回归任务,RF 的最终输出为所有决策树模拟结果值的算数平均值:

$$\hat{y} = \frac{1}{M} \sum_{m=1}^M f_m(X) \quad (1)$$

其中, X 为特征向量, $f_m(X)$ 表示第 m 棵决策树的模拟结果值。

2.1.2 模型构建与变量组合设计 本研究针对不同预报时效(1/3/7 d),分别构建三个独立的 RF 模型。具体的蓝藻水华面积预报流程如图 3(c)所示,主要包括以下四个步骤:

(1) **数据集匹配**:针对每一预报时效(1/3/7 d),模型的输入特征(X)为历史最佳累积环境变量(提取过程详见 1.2.3 及 1.2.4 节)。目标变量(y)为目标预报日期(t)的蓝藻水华面积(经 \log_{10} 变换)。通过将特征矩阵与目标变量对齐,分别构建对应于三个预报时效的数据集 D_{t-1} , D_{t-3} , D_{t-7} 。

(2) **变量组合穷举与变量间共线性诊断**:鉴于蓝藻水华与环境变量之间存在非线性关系及交互作用,纳入所有可能的变量组合有助于揭示这些潜在的相互关系。因此,本研究采用穷举法^[40]生成所有可能的非空变量组合。对于包含 n 个候选变量的特征集,其非空子集总数为 $2^n - 1$,即对应 $2^n - 1$ 种变量组合方案。此外,为评估变量间多重共线性对模型稳定性的潜在影响,对所有变量进行共线性诊断,具体包括计算各变量的方差膨胀因子(Variance Inflation Factor, VIF)与容忍度(Tolerance, TOL):

$$VIF_j = \frac{1}{1 - R_j^2} \quad (2)$$

$$TOL_j = 1 - R_j^2 \quad (3)$$

其中, R_j^2 为变量 j 对其余变量回归的决定系数。仅保留 $VIF < 5$ 且 $TOL > 0.2$ 的低共线性的变量进入模型训练。

(3) **模型训练与评估**:针对每一个通过共线性检验的变量组合,考虑到数据集样本量相对有限(共 933 个样本),若采用常规的随机划分方法(如 7:3 划分训练集与测试集)可能导致训练样本不足,且单次划分结果对模型评估的波动性较大,进而影响模型稳定性和泛化能力。为此,本研究采用留一法交叉验证(Leave-One-Out Cross Validation, LOOCV)进行模型训练与性能评估。具体而言,在每次迭代中,从数据集 $D \in \{D_{t-1}, D_{t-3}, D_{t-7}\}$ 中保留一个样本作为测试集,其余 932 个样本作为训练集用于构建 RF 模型。该过程重复执行 933 次,确保每个样本均被独立地用于测试。最终,模型的整体性能基于所有交叉验证迭代的测试结果汇总得出。此外,模型性能评估指标包括决定系数(R^2)以及均方根误差(Root Mean Square Error,

RMSE):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (5)$$

其中, y_i 为蓝藻水华面积的真实值, \hat{y}_i 为蓝藻水华面积的预报值, \bar{y} 为蓝藻水华面积真实值的平均值, n 为 D_{test} 样本数量。

(4) 最优变量组合选择: 比较所有变量组合 (共 2^n-1 种) 在测试集上的 R^2 和 RMSE 表现, 并综合考虑模型的简洁性 (即变量个数), 最终为每一预报时效 (1/3/7 d) 选择一个最优变量组合, 用于最终的蓝藻水华面积预报。

2.1.3 对比实验 为评估时间累积变量对预报精度的影响, 本研究增设对比实验: 分别采用第 $t-1$ d、第 $t-3$ d 和第 $t-7$ d 的单日变量 (表 1) 构建模型, 与上述基于累积变量的模型进行性能比较。

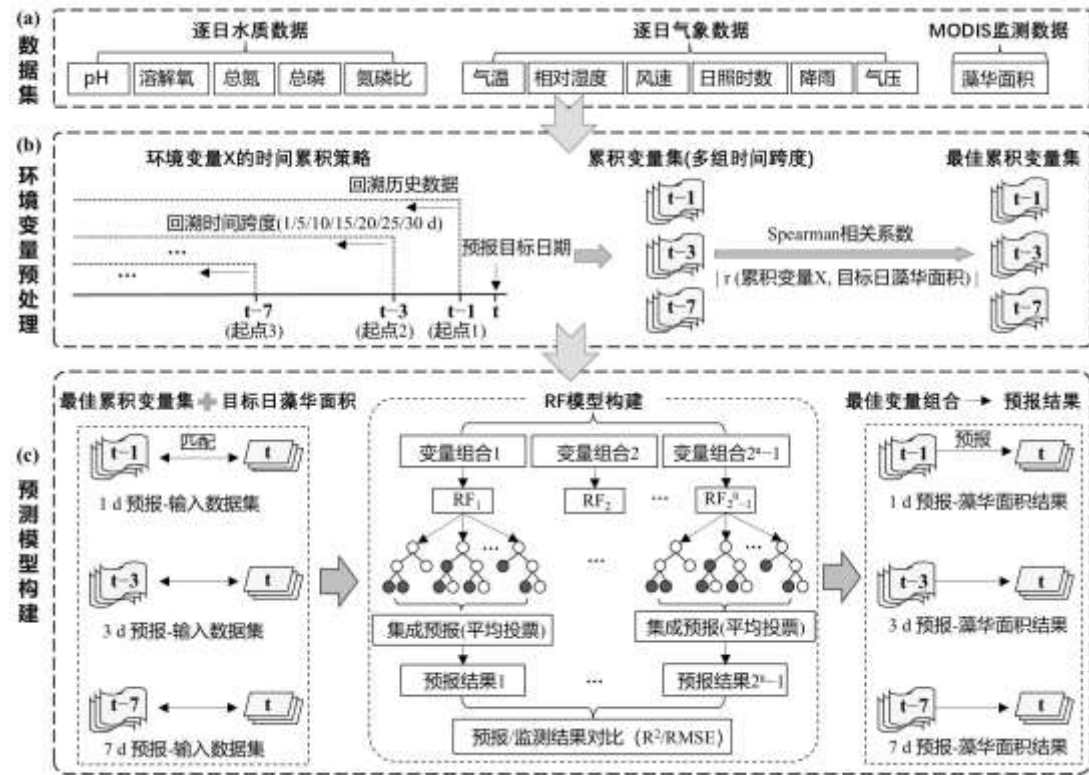


图 3 基于时间累积变量及 RF 模型的蓝藻水华面积预测流程

Fig.3. Prediction workflow of cyanobacterial bloom area based on accumulation variables and the RF model

2.2 基于 SHAP 的模型可解释性

为解析 RF 模型的决策机制, 本研究采用 SHAP 方法进行可解释性分析。该方法源于博弈论中的 Shapley 值理论, 通过计算每个特征在所有特征组合中的边际贡献均值, 实现对 RF 模型预测结果的量化解释^[41]。与传统特征重要性排序方法相比, SHAP 方法具备以下特点: (1) 公平性: 依据特征的实际影响分配其贡献值; (2) 一致性: 保持特征重要性与模型输出的逻辑关联; (3) 可加性: 各特征贡献值之和等于模型预测值与实际值之差。其核心公式如下:

$$\phi_i(f) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f(S \cup \{i\}) - f(S)] \quad (6)$$

其中， $\phi_i(f)$ 为特征*i*的 Shapley 值，表示特征*i*对预报结果的贡献；*N*为所有特征的集合；*S*为特征集合，表示除了特征*i*以外的特征； $f(S)$ 为模型在特征集合*S*上的预报值； $f(S \cup \{i\})$ 为模型在特征集合*S* ∪ {*i*}上的预报值；|*S*|和|*N*|分别是集合*S*和集合*N*的特征个数。

为识别关键环境变量及其促进蓝藻水华的阈值范围。本研究采用两种 SHAP 可视化方法：（1）SHAP 概要图，用于从全局角度展示各变量的总体贡献度及影响方向；（2）SHAP 依赖图，用于从局部尺度解析单一变量与模型预报结果之间的非线性关系。

2.3 语言实现

本研究的数据处理、累积变量构建、机器学习建模与验证以及 SHAP 可视化分析均在 Python (v3.8) 环境下完成。具体地，使用 scikit-learn(v1.3.2) 库构建 RF 模型及 LOOCV 法验证，使用 shap(v0.44.1) 库执行模型可解释性分析。

3 结果

3.1 变量的时间累积效应分析以及共线性诊断

环境变量对蓝藻水华的影响表现出时间累积效应（图 4）。分析结果显示，单日变量（累积时间=1 d）与蓝藻水华的相关性整体偏低，而在延长累积时间后，多数变量与蓝藻水华面积的相关性增强。这一结果说明，仅依赖单日变量可能难以充分反映变量对蓝藻水华的影响。表 2 总结了各变量的最佳累积时间，即其与蓝藻水华面积的绝对相关系数达到最大时所对应的累积天数。其中，平均风速以及日照时数与蓝藻水华面积的相关性较弱（ $r < 0.2$ ），在后续分析中未予纳入。最大风速对蓝藻水华影响的最佳累积时间为 1d，未表现出长期累积效应。气温、相对湿度、降雨和气压等气象变量的最佳累积时间多在 15~30 d 之间，这与其通过影响水体分层和混合过程间接作用于蓝藻水华的机制相符^[42, 43]。水质变量的最佳累积时间普遍较短，多为 1~10 d，反映出其对蓝藻水华生长具有更为直接的短期响应关系。

表 2 环境变量的最佳累积时间及对应的相关系数¹⁾

Table 1. Optimal accumulation time windows for environmental variables and corresponding correlation coefficients.

	1 d 预报		3 d 预报		7 d 预报	
	最佳累积时间长度 (d)	r	最佳累积时间长度 (d)	r	最佳累积时间长度 (d)	r
pH	25	0.64**	25	0.61**	20	0.60**
溶解氧	10	-0.62**	10	-0.60**	5	-0.52**
总磷	10	0.45**	5	0.44**	5	0.41**
总氮	5	0.47**	5	0.46**	1	0.44**
氮磷比	10	-0.68**	5	-0.66**	5	-0.61**
平均气温	30	0.66**	20	0.65**	20	0.64**
最高气温	30	0.66**	20	0.64**	20	0.63**
最低气温	30	0.66**	20	0.65**	20	0.63**
相对湿度	20	0.36**	15	0.32*	15	0.29*
平均风速	1	-0.08*	1	0.07*	1	0.07*
最大风速	1	-0.43**	1	-0.21**	1	-0.12**
日照时数	30	0.18**	30	0.19**	25	0.19**
降雨量	30	0.34**	30	0.32**	30	0.31**
气压	30	-0.61**	30	-0.61**	30	-0.59**

1) 相关系数 *r* 中，**表示该结果通过了显著性检验 ($p < 0.01$)，*表示该结果通过了更严格的显著性检验 ($p < 0.05$)。

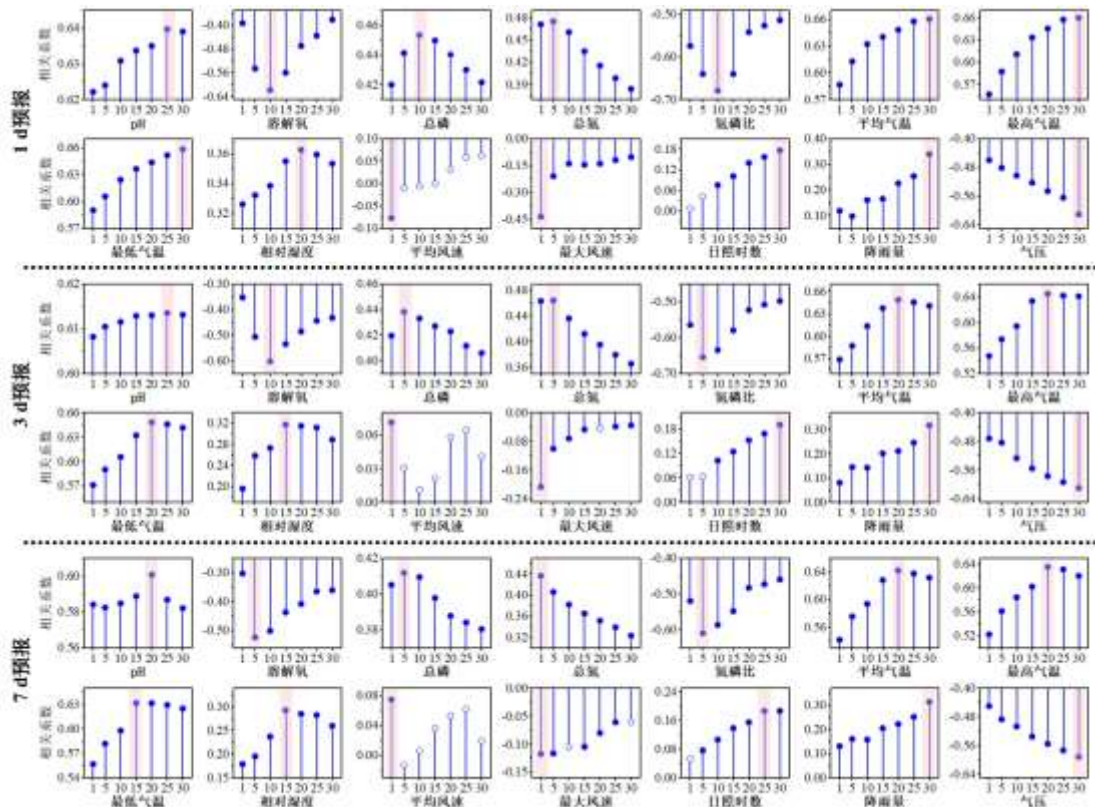


图4 环境变量在不同累积时间长度下与蓝藻水华面积的相关关系 (“实心”点表示通过显著性检验 ($p < 0.05$), “空心”点表示未通过显著性检验, 粉色阴影标记了每个环境变量的最佳累积时间所在位置)

Fig. 4. Correlation patterns between environmental variables and cyanobacterial bloom area under different temporal accumulation lengths (solid markers indicate statistically significant relationships ($p < 0.05$), hollow markers denote non-significant correlations. Pink shading marks the optimal cumulative time position for each environmental variable).

基于相关性分析结果, 共筛选出 8 个与蓝藻水华面积具有较强相关性的变量: pH、溶解氧、氮磷比、平均气温、最大风速、相对湿度、降雨量和气压。其中, 平均气温 ($r \geq 0.64$) 和氮磷比 ($r \leq -0.61$) 分别为温度系列和氮磷系列中与蓝藻水华面积相关性最高的指标。通过方差膨胀因子(VIF)和容差(TOL)进行共线性诊断(图5), 结果显示, 在 1/3/7 d 预报中, 平均气温与气压之间存在共线性 ($VIF > 5, TOL < 0.2$)。由于气压与蓝藻水华面积的相关性低于平均气温, 因此在建模中予以剔除。最终保留的 7 个变量(pH、溶解氧、氮磷比、平均气温、最大风速、相对湿度、降雨量)均通过共线性检验, 并用于后续 RF 模型的最优变量组合训练。

3.2 最佳变量组合选择结果

为比较单日变量与时间累积变量在预报性能上的差异, 本研究对两种特征构建方法分别进行了最优特征组合搜索。如图 6 所示, 变量组合的选择对预报精度具有明显影响, 且预报性能并非随特征数量增加而持续提升, 而是存在峰值区间。具体而言, 在使用累积变量进行预报时, 1 d 预报在 6 个特征时达到最佳性能 ($R^2 = 0.75$), 3 d 预报在 4 个特征时 R^2 最高 (0.69), 7 d 预报在 5 个特征时获得 0.66 的 R^2 值。基于单日变量的预报同样通过穷尽搜索获得各特征数量下的最优性能(图 6a)。尽管两类变量在最优组合的构成上有所不同, 但无论特征数量如何变化, 在所有预报时效中, 累积变量组合的预报性能 ($R^2: 0.40\sim 0.75$) 均优于单日变量组合 ($R^2: 0.37\sim 0.63$), 说明引入时间累积信息构建特征能够提升模型预报精度的上限。

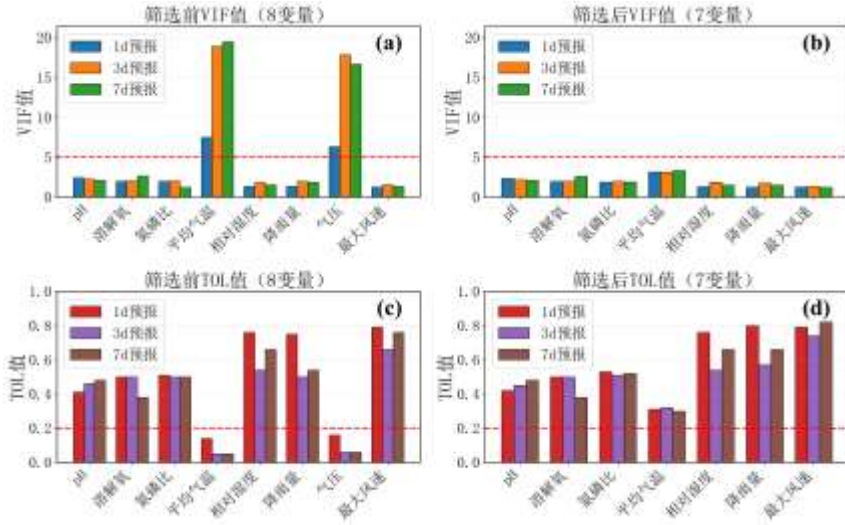


图5 变量之间的共线性诊断结果 (诊断阈值设定为 $VIF > 5$, $TOL < 0.2$, 以红色阈值线标出)。
 Fig.5. Diagnosis results for multicollinearity among variables (diagnostic thresholds set at $VIF > 5$ and $TOL < 0.2$, indicated by red threshold lines).

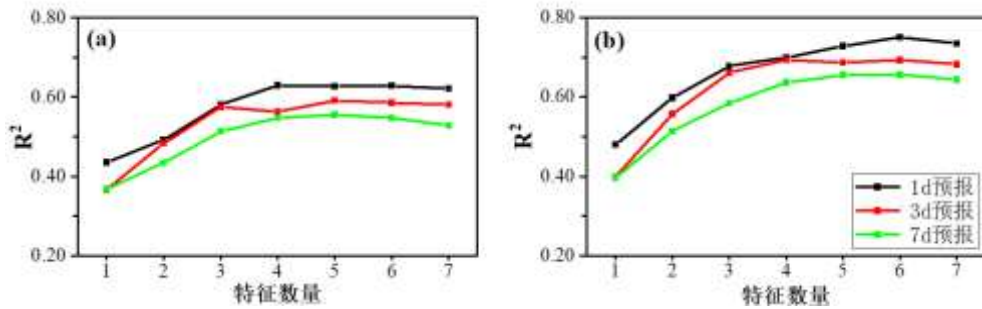


图6 不同特征数量下最优组合的最优精度比较: (a) 基于单日变量的预报; (b) 基于时间累积变量的预报
 图中展示了各自特征构建策略下, 针对每个特征数量找到的全局最优组合的性能

Fig.6 Accuracy comparison of optimal variable combinations with different feature numbers: Prediction based on (a) single-day variables; and (b) accumulation variables. The figure illustrates the performance of the globally optimal combination identified for each variable quantity under their respective feature construction strategies.

表3列出了不同预报时效下的最优累积变量组合。结果显示, pH、溶解氧、平均气温和降雨量在所有预报时效中均被选取为最优组合的变量, 表明这4个指标在蓝藻水华预报中具有普遍适用性与稳健性。此外, 氨磷比在1d和7d预报的最优组合中均被纳入, 表明其在这两个预报时效中具有一定的作用。此外, 最大风速仅出现在1d预报的最优组合中, 说明其对蓝藻水华的影响不具备长期的时滞累积效应, 前1d的最大风速对次日藻华发生具有相对较强的指示意义。

3.3 蓝藻水华面积预报结果

本研究对1/3/7d不同预报时效的模型性能进行了评估(表4)。结果显示, 基于时间累积变量的预报模型在各项评估指标上均优于单日变量模型。具体而言, 1d预报表现最佳($R^2 = 0.75$, $RMSE = 49.37 \text{ km}^2$), 其预报精度较3d预报($R^2 = 0.69$, $RMSE = 55.43 \text{ km}^2$)和7d预报($R^2 = 0.66$, $RMSE = 57.56 \text{ km}^2$)相比, R^2 分别高出了8.70%和13.64%, $RMSE$ 分别降低了10.93%和14.23%。

不同时效的预报结果与MODIS观测结果的对比表明(图7), 基于累积变量的预报值与观测值之间的

一致性优于基于单日变量的模型，在预报大规模蓝藻水华事件（蓝藻水华面积占比 >30%）时，这一优势表现明显。具体而言，单日变量模型普遍低估大规模藻华面积，累积变量模型虽随预报时效的延长也呈现一定低估趋势，但其偏差程度相对较低。说明累积变量能够捕捉蓝藻水华暴发的滞后响应机制，改善模型的预报稳健性与精度。

表 3 不同预报时效下的最佳变量组合¹⁾

Tab.3. Optimal variable combinations for different forecast lead times.

预报时效	最优累积变量组合
1 d	pH(25) + 溶解氧(10) + 氮磷比(10) + 平均气温(30) + 降雨(30) + 最大风速(1)
3 d	pH(25) + 溶解氧(10) + 平均气温(20) + 降雨(30)
7 d	pH(20) + 溶解氧(5) + 氮磷比(5) + 平均气温(20) + 降雨(30)

1) “()”中的数字表示最佳累积时间长度。

表 4 不同预报时效下的训练集和验证集精度。

Tab.4 Accuracy of training and validation sets for different forecast lead times.

预报时效	时间累积变量				单日变量			
	训练集 RMSE	验证集 RMSE	训练集 R ²	验证集 R ²	训练集 RMSE	验证集 RMSE	训练集 R ²	验证集 R ²
1 d	33.90	49.37	0.88	0.75	49.27	60.26	0.75	0.63
3 d	42.33	55.43	0.82	0.69	53.97	63.32	0.70	0.59
7 d	43.35	57.56	0.81	0.66	54.62	65.52	0.70	0.56

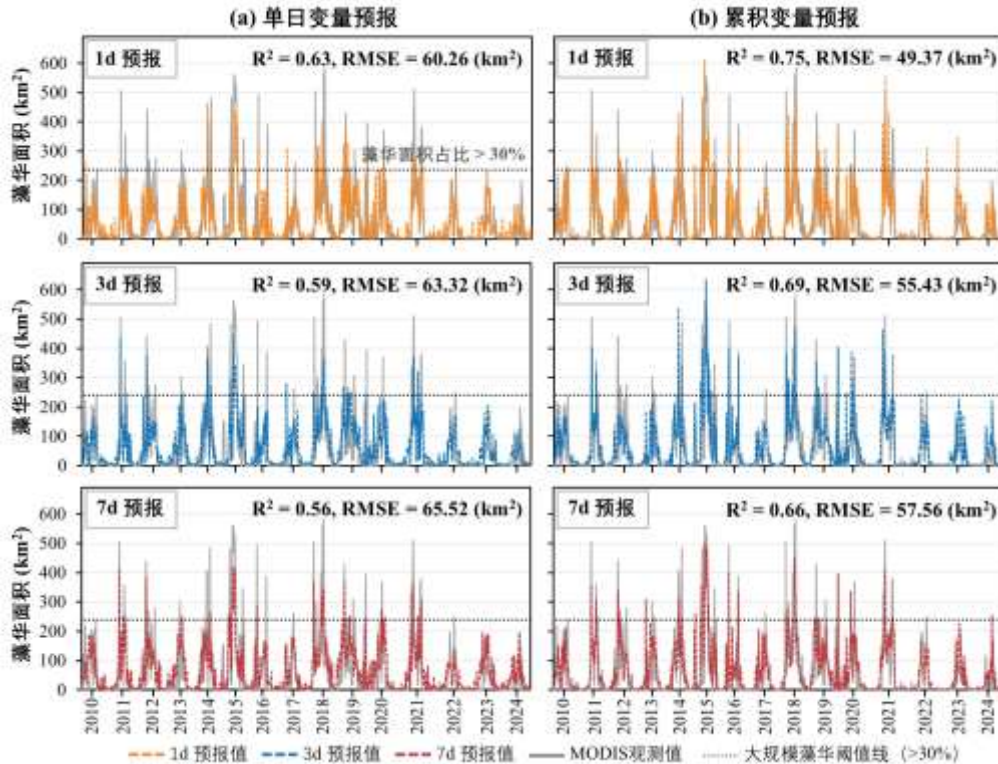


图 7 不同预报时效下的蓝藻水华面积预报结果：(a) 基于单日变量的预报；(b) 基于时间累积变量的预报

Fig.7 Prediction results of cyanobacterial bloom area for different forecast lead times: Prediction based on (a) single-day variables, and (b) accumulation variables.

3.4 蓝藻水华面积扩张的关键因素

SHAP 概述图 (图 8a) 展示了不同预报时效下环境变量的贡献程度 (按 SHAP 值降序排列)。平均气温在不同的预报时效中均具有重要影响。在 1 d 预报中, 氮磷比和平均气温的贡献度最高; 3 d 与 7 d 预报则以平均气温和 pH 为主要影响因子。这些变量在的 SHAP 值分布范围较宽, 反映其在蓝藻水华形成过程中具有持续影响。

关系依赖图 (图 8b) 展示了各变量实际监测值与对应 SHAP 值的关系 (SHAP 值大于 0 表示该变量对蓝藻水华发生具有正向促进作用)。分析显示, 变量对蓝藻水华的促进作用存在阈值特征: 当平均气温约 $> 23^{\circ}\text{C}$ 、最大风速约 $< 4\text{ m/s}$ 、降雨量约 $> 200\text{ mm}$ 、氮磷比约 < 15 、pH 约 > 8.5 、溶解氧约 $< 8.9\text{ mg/L}$ 时, SHAP 值转为正向。并且该促进作用呈现非线性特征: 当平均气温约 $> 33^{\circ}\text{C}$ 、降雨量约 $> 400\text{ mm}$ 、pH 约 > 9.0 时, 促进作用减弱并逐渐转为抑制, 表明极端环境条件可能改变蓝藻的生长动态。

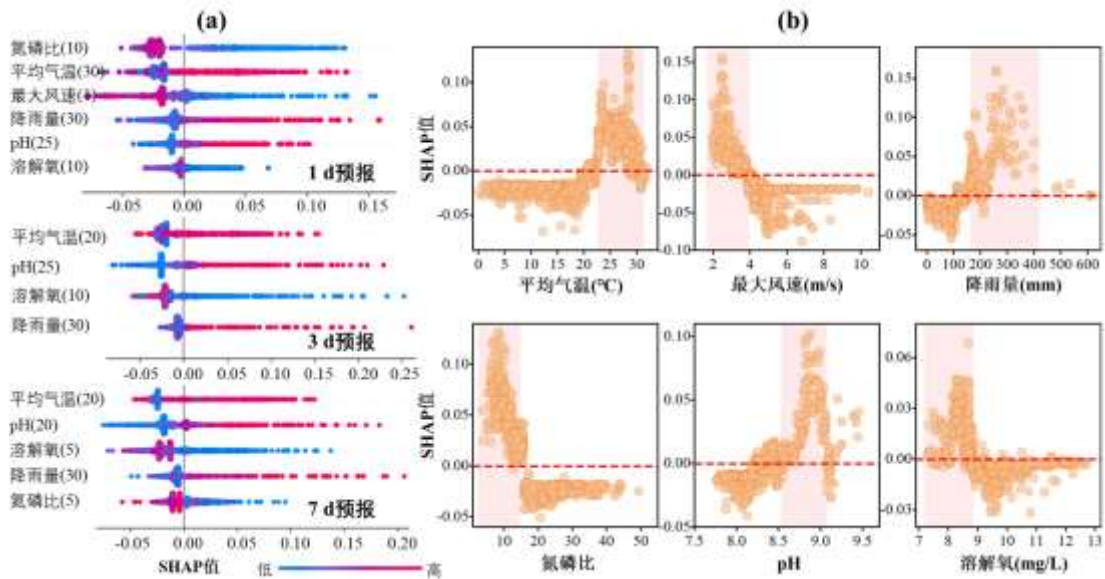


图 8 变量重要性与非线性依赖关系。(a) 为 SHAP 概述图 (图中散点代表每个样本的 SHAP 值)。(b) 为 SHAP 关系依赖图 (由于不同组合中的变量存在重复, 重复变量仅展示 1 d 预报中的变量)。粉色阴影为促进蓝藻水华发生的变量区间

Fig.8 Variable importance and nonlinear dependencies in cyanobacterial bloom prediction. (a) SHAP summary plot (Scatter points represent SHAP values for each sample). (b) SHAP dependence plot (Focuses on variables from the 1-day lead time prediction due to duplicate variables in different combinations). The pink shading indicates the variable range promoting cyanobacterial blooms.

4 结论与讨论

本研究以中国富营养化湖泊巢湖为研究对象, 基于 2010–2024 年多站点气象、水质变量及 MODIS 卫星数据, 分析了各变量与蓝藻水华在 1–30 d 时间尺度上的累积效应, 识别了各变量的最佳累积时间并构建最优变量组合。利用 RF 模型, 分别实现了提前 1 d、3 d 和 7 d 的蓝藻水华面积预报, 并结合 SHAP 方法解析了关键变量对蓝藻水华的影响及促进阈值区间。主要结论与讨论如下:

蓝藻水华的暴发是环境变量逐步累积作用的结果, 而非瞬时现象^[44]。本研究结果显示, 多数环境变量在考虑时间累积效应后, 其与蓝藻水华面积的相关性明显增强 (图 4)。并且引入时间累积变量后, 模型预报精度得到提升, 其中 1d 预报的 R^2 提高了 19.05% (RMSE 降低 18.07%), 3 d 预报提高了 16.95% (RMSE 降低 12.46%), 7 d 预报提高了 17.86% (RMSE 降低 12.15%)。表明采用变量累积策略能够捕捉环境变化的滞后效应, 增强模型对蓝藻水华发生的预报能力。其中, 对于气象累积变量, 持续较高的温度有助于促进蓝藻的代谢与繁殖。当 30d 平均气温高于约 23°C 时, 对蓝藻水华形成表现为促进作用; 而当温度超过约 33°C 后, 该作用减弱 (图 8b)。相关研究指出, 温度过高可能会抑制光合酶活性, 从而减弱蓝藻水华的

规模^[45]。最大风速仅在 1 d 预报中具有明显影响, 风速低于 4 m/s 左右时有利于蓝藻水华形成, 而 3 d 与 7 d 预报中风速未表现出显著贡献, 反映风速对蓝藻水华的影响主要为即时作用, 缺乏长期累积效应。当降雨总量超过约 200 mm 时, 其对蓝藻水华呈现促进作用, 可能与降雨引起的地表径流输入外源营养物质有关; 然而, 当降雨量增至约 400 mm 以上时, 该促进效应减弱并转为抑制, 可能是由于强降雨引起的水体混合与湍流, 扰乱了藻细胞表层聚集, 从而阻碍蓝藻水华形成。与气象变量相比, 氮、磷及溶解氧等水质指标对蓝藻水华的影响表现出较短的累积时间(通常在 10d 以内)。当氮磷比约低于 15、溶解氧约低于 8.9 mg/L 时, 对蓝藻水华具有促进作用。pH 对蓝藻水华的影响具有较长的累积时间(约 25d), 当 pH 高于 8.5 时, 碱性环境有利于提升蓝藻的光合作用效率, 为其生长创造有利条件。因此, 这些关键区间在蓝藻水华的预报和防控过程中应给予特别重视。但应认识到, 本研究基于现有数据集得出的预报性能与关键阈值区间, 在未来拓展观测年份或增加监测点位后可能存在优化空间。这一不确定性源于数据驱动模型的普遍特性: 模型性能及结果的稳健性依赖于训练样本的规模与代表性。构建长期、全面的数据集是提升模型普适性、降低结论不确定性的关键途径。

对于研究目标, 以往研究多采用分类模型预测蓝藻水华在未来是否发生或发生的严重等级^[25, 46], 或者以叶绿素 a 浓度、藻类生物量和微囊藻毒素浓度等作为预报对象^[47-49]。这些指标虽能反映蓝藻的生化状态, 但难以直观体现蓝藻水华的覆盖面积。相比之下, 本研究以蓝藻水华面积为预报目标, 有助于管理者更直观地评估灾害的实际发生规模, 为快速部署防控资源、应对大规模藻华事件提供依据。

然而, 本研究仍存在一定的局限性。首先, 在变量选择方面, 某些预测变量(如 pH 和溶解氧)的影响主要基于统计相关性, 而非明确的生态生理机制。尽管这些变量有助于提升模型预测精度, 但其与蓝藻水华之间的因果关系及直接作用路径仍需进一步验证。同样, pH 和溶解氧作为预测变量虽通过历史累积值表征了环境遗留效应, 但这种反馈机制的具体路径仍需结合控制实验或过程机理模型进行探讨。其次, 本研究的预报任务完全依赖历史环境数据。已有研究利用遥感数据与未来气象信息来预报蓝藻水华^[34, 50]。本研究采用的历史数据外推策略虽可避免未来气象数据的不确定性, 但未引入天气预报信息, 可能未充分考虑蓝藻水华发生当日的瞬时气象条件(如风速、风向)的影响。然而, 天气预报数据通常覆盖范围较广, 单纯依赖该数据可能难以准确反映湖泊局地的具体状况。此外, 当前模型仅针对蓝藻水华暴发面积的时间变化进行预报, 未涉及其空间分布信息(如饮用水取水口、旅游区等敏感区域), 但所设计的变量累积和组合策略仍可为未来空间分布预报提供技术参考。最后, 蓝藻水华的发生与消落过程常伴随较强的时空异质性。尤其在夏秋季节, 蓝藻水华的突发与消退常受多因子协同影响(如温度骤变、短时强降雨、水动力扰动等)。由于未引入小时尺度气象数据或高频水质监测信息, 模型在捕捉蓝藻水华快速转变状态时可能存在一定偏差。

未来研究应从多源数据融合与模型集成这两个方向推进:(1)联合历史气象数据与精细化天气预报(如 ECMWF 高分辨率数值模型), 尤其是引入小时尺度的温度、风速等气象变量, 以提升预报精度;(2)构建机理模型与数据驱动模型的混合架构, 以弥补纯数据驱动模型在极端事件和突变情景下的预报偏差;(3)引入时空预报技术, 如结合卷积神经网络(Convolutional Neural Network, CNN)、图卷积神经网络(Graph Convolutional Network, GCN)、遥感技术、空间插值或分区建模等方法, 实现蓝藻水华的空间分布预报。此外, 在模型可移植性方面, 针对监测数据稀缺的湖泊, 以及 MODIS 卫星即将退役的实际情况, 未来研究应更加关注数据源的替代性和灵活性。可以优先整合多源遥感数据(Sentinel-2/Landsat 8/GOCI-II 等), 以保障蓝藻水华长期连续预报的数据连续性。对于复杂异质的水体环境, 可综合考虑多源遥感、精细化环境变量、机制-数据驱动混合建模, 或通过分区训练(如按湖区、水深、营养盐梯度划分)等方式, 提升模型在不同场景下的泛化能力。

致谢: 感谢南京水利科学研究院生态环境研究所马金戈博士提供的遥感影像数据支持。

5 参考文献

[1] Liu D., H. Duan, S. Loiselle *et al.* Observations of water transparency in China's lakes from space. *International Journal of Applied*

- Earth Observation and Geoinformation*, 2020, **92**: 102187. DOI: 10.1016/j.jag.2020.102187.
- [2] Dai Y., S. Yang, D. Zhao *et al.* Coastal phytoplankton blooms expand and intensify in the 21st century. *Nature*, 2023, **615**(7951): 280-284. DOI: 10.1038/s41586-023-05760-y.
- [3] Ho J.C., A.M. Michalak, N. Pahlevan. Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature*, 2019, **574**(7780): 667-670. DOI: 10.1038/s41586-019-1648-7.
- [4] Luo J., H. Duan, Y. Xu *et al.* Global trends and regime state shifts of lacustrine aquatic vegetation. *The Innovation*, 2025, **6**(3): 100784. DOI: 10.1016/j.xinn.2024.100784.
- [5] Luo J., G. Ni, Y. Zhang *et al.* A new technique for quantifying algal bloom, floating/emergent and submerged vegetation in eutrophic shallow lakes using Landsat imagery. *Remote Sensing of Environment*, 2023, **287**: 113480. DOI: 10.1016/j.rse.2023.113480.
- [6] Kong FX, Gao G. Hypothesis on cyanobacteria bloom-forming mechanism in large shallow eutrophic lakes. *Acta Ecologica Sinica*, 2005, **25**(03): 589-595. DOI: 10.3321/j.issn:1000-0933.2005.03.028. [孔繁翔, 高光. 大型浅水富营养化湖泊中蓝藻水华形成机理的思考. *生态学报*, 2005, 25(3): 589-595.]
- [7] Carmichael W.W., S.M.F.O. Azevedo, J.S. An *et al.* Human Fatalities from Cyanobacteria: Chemical and Biological Evidence for Cyanotoxins. *Environmental Health Perspectives*, 2001, **109**(7): 663-668. DOI: 10.1289/ehp.01109663.
- [8] Qin B., G. Zhu, G. Gao *et al.* A Drinking Water Crisis in Lake Taihu, China: Linkage to Climatic Variability and Lake Management. *Environmental Management*, 2009, **45**(1): 105-112. DOI: 10.1007/s00267-009-9393-6.
- [9] Steffen M.M., T.W. Davis, R.M.L. McKay *et al.* Ecophysiological Examination of the Lake Erie Microcystis Bloom in 2014: Linkages between Biology and the Water Supply Shutdown of Toledo, OH. *Environmental Science & Technology*, 2017, **51**(12): 6745-6755. DOI: 10.1021/acs.est.7b00856.
- [10] Wynne T., R. Stumpf. Spatial and Temporal Patterns in the Seasonal Distribution of Toxic Cyanobacteria in Western Lake Erie from 2002–2014. *Toxins*, 2015, **7**(5): 1649-1663. DOI: 10.3390/toxins7051649.
- [11] Wang H., C. Xu, Y. Liu *et al.* From unusual suspect to serial killer: Cyanotoxins boosted by climate change may jeopardize megafauna. *The Innovation*, 2021, **2**(2): 100092. DOI: 10.1016/j.xinn.2021.100092.
- [12] Paerl H.W., J. Huisman. Blooms like it hot. *Science*, 2008, **320**(5872): 57-58. DOI: 10.1126/science.1155398.
- [13] Tigli M., M.P. Bak, J.H. Janse *et al.* The future of algal blooms in lakes globally is in our hands. *Water Research*, 2024, **268**: 122533. DOI: 10.1016/j.watres.2024.122533.
- [14] Ho J.C., A.M. Michalak. Phytoplankton blooms in Lake Erie impacted by both long-term and springtime phosphorus loading. *Journal of Great Lakes Research*, 2017, **43**(3): 221-228. DOI: 10.1016/j.jglr.2017.04.001.
- [15] Qin M., Z. Li, Z. Du. Red tide time series forecasting by combining ARIMA and deep belief network. *Knowledge-Based Systems*, 2017, **125**: 39-52. DOI: 10.1016/j.knosys.2017.03.027.
- [16] Stumpf R.P., L.T. Johnson, T.T. Wynne *et al.* Forecasting annual cyanobacterial bloom biomass to inform management decisions in Lake Erie. *Journal of Great Lakes Research*, 2016, **42**(6): 1174-1183. DOI: 10.1016/j.jglr.2016.08.006.
- [17] Kim J., T. Lee, D. Seo. Algal bloom prediction of the lower Han River, Korea using the EFDC hydrodynamic and water quality model. *Ecological Modelling*, 2017, **366**: 27-36. DOI: 10.1016/j.ecolmodel.2017.10.015.
- [18] Tang TJ, Yang S, Yin KH *et al.* Simulation of eutrophication in Shenzhen Reservoir based on EFDC model. *Journal of Lake Sciences*, 2014, **26**(3): 393-400. DOI: 10.3969/j.issn.1003-5427.2014.03.009. [唐天均, 杨晟, 尹魁浩等. 基于 EFDC 模型的深圳水库富营养化模拟. *湖泊科学*, 2014, 26(3): 393-400.]
- [19] Zia A., A. Bombliès, A.W. Schroth *et al.* Coupled impacts of climate and land use change across a river–lake continuum: insights from an integrated assessment model of Lake Champlain’s Missisquoi Basin, 2000–2040. *Environmental Research Letters*, 2016, **11**(11): 114026. DOI: 10.1088/1748-9326/11/11/114026.
- [20] Gupta A., M.M. Hantush, R.S. Govindaraju. Sub-monthly time scale forecasting of harmful algal blooms intensity in Lake Erie using remote sensing and machine learning. *Science of The Total Environment*, 2023, **900**: 165781. DOI: 10.1016/j.scitotenv.2023.165781.
- [21] Kim D., J.-H. Lim, Y. Chun *et al.* Phytoplankton nutrient use and CO₂ dynamics responding to long-term changes in riverine N and P availability. *Water Research*, 2021, **203**: 117510. DOI: 10.1016/j.watres.2021.117510.

- [22] Villanueva P., J. Yang, L. Radmer *et al.* One-Week-Ahead Prediction of Cyanobacterial Harmful Algal Blooms in Iowa Lakes. *Environmental Science & Technology*, 2023, **57**(49): 20636-20646. DOI: 10.1021/acs.est.3c07764.
- [23] Xu H, Dai CR, He YQ *et al.* Quantitative assessment and prediction of the effects of meteorological conditions on the occurrence of cyanobacteria bloom in Dianchi Lake based on random forest. *Journal of Hydroecology*, 2024: 1-8. DOI: 10.15928/j.1674-3075.202308040210. [徐虹, 戴丛蕊, 何雨岑等. 基于随机森林定量评估气象条件对滇池蓝藻水华发生的影响及预测. 水生生态学杂志, 2024, 1-8.]
- [24] Li YM, Tan ZY, yang C *et al.* Extraction of Algal Blooms in Dianchi Lake Based on Multi-Source Satellites Using Machine Learning Algorithms. *Advances in Earth Science*, 2022, **37**(11): 1141-1156. DOI: 10.11867/j.issn.1001-8166.2022.064. [李一民, 谭振宇, 杨辰等. 基于多源卫星的滇池藻华提取机器学习算法研究. 地球科学进展, 2022, 37(11): 1141-1156.]
- [25] Ai H., K. Zhang, J. Sun *et al.* Short-term Lake Erie algal bloom prediction by classification and regression models. *Water Research*, 2023, **232**: 119710. DOI: 10.1016/j.watres.2023.119710.
- [26] Song K., C. Fang, P.-A. Jacinthe *et al.* Climatic versus Anthropogenic Controls of Decadal Trends (1983–2017) in Algal Blooms in Lakes and Reservoirs across China. *Environmental Science & Technology*, 2021, **55**(5): 2929-2938. DOI: 10.1021/acs.est.0c06480.
- [27] Wang Q., L. Sun, Y. Zhu *et al.* Hysteresis effects of meteorological variation-induced algal blooms: A case study based on satellite-observed data from Dianchi Lake, China (1988–2020). *Science of The Total Environment*, 2022, **812**: 152558. DOI: 10.1016/j.scitotenv.2021.152558.
- [28] Zhao XB, Zhang Q, Yang FS *et al.* Analysis of Influencing Factors of PM2.5 in Shanxi Province Based on XGBoost-SHAP Method. *Research of Environmental Sciences*, 2025: 1-16. DOI: 10.13198/j.issn.1001-6929.2025.03.16. [赵兴赞, 张强, 杨方社等. 基于XGBoost-SHAP方法的陕西省PM2.5影响因素分析. 环境科学研究, 2025, 1-16.]
- [29] He B., X. Zhu, Z. Cang *et al.* Interpretation and Prediction of the CO2 Sequestration of Steel Slag by Machine Learning. *Environmental Science & Technology*, 2023, **57**(46): 17940-17949. DOI: 10.1021/acs.est.2c06133.
- [30] Chen P., J. Luo, Z. Xiong *et al.* Can the establishment of a protected area improve the lacustrine environment? A case study of Lake Chaohu, China. *Journal of Environmental Management*, 2023, **342**: 118152. DOI: 10.1016/j.jenvman.2023.118152.
- [31] Zhou Z., Y. Liu, S. Wang *et al.* Interactions between Phosphorus Enrichment and Nitrification Accelerate Relative Nitrogen Deficiency during Cyanobacterial Blooms in a Large Shallow Eutrophic Lake. *Environmental Science & Technology*, 2023, **57**(7): 2992-3001. DOI: 10.1021/acs.est.2c07599.
- [32] Ma J., S. Loiselle, Z. Cao *et al.* Unbalanced impacts of nature and nurture factors on the phenology, area and intensity of algal blooms in global large lakes: MODIS observations. *Science of The Total Environment*, 2023, **880**: 163376. DOI: 10.1016/j.scitotenv.2023.163376.
- [33] Kim K., H. Mun, H. Shin *et al.* Nitrogen Stimulates Microcystis-Dominated Blooms More than Phosphorus in River Conditions That Favor Non-Nitrogen-Fixing Genera. *Environmental Science & Technology*, 2020, **54**(12): 7185-7193. DOI: 10.1021/acs.est.9b07528.
- [34] Du Y., H. Zhao, J. Li *et al.* Cyanobacterial blooms prediction in China's large hypereutrophic lakes based on MODIS observations and Bayesian theory. *Journal of Hazardous Materials*, 2024, **480**: 136057. DOI: 10.1016/j.jhazmat.2024.136057.
- [35] Miura Y., H. Imamoto, Y. Asada *et al.* Prediction of algal bloom using a combination of sparse modeling and a machine learning algorithm: Automatic relevance determination and support vector machine. *Ecological Informatics*, 2023, **78**: 102337. DOI: 10.1016/j.ecoinf.2023.102337.
- [36] Biau G. Analysis of a Random Forests Model. *Journal of Machine Learning Research*, 2010, **13**(2): 1063-1095. DOI: 10.1109/TASE.2012.2183739.
- [37] Li XY, Wang H, Wu XM *et al.* Characterization and prediction of dissolved oxygen fluctuation in Poyang Lake based on machine learning. *Journal of Lake Sciences*, 2025, **37**(3):915-927. DOI: 10.18307/2025.0328. [李晓瑛, 王华, 吴小毛等. 基于机器学习的鄱阳湖溶解氧波动特征及预测. 湖泊科学, 2025, 37(3):915-927.]
- [38] Saravani M.J., R. Noori, C. Jun *et al.* Predicting Chlorophyll-a Concentrations in the World's Largest Lakes Using Kolmogorov-Arnold Networks. *Environmental Science & Technology*, 2025, **59**(3): 1801-1810. DOI: 10.1021/acs.est.4c11113.
- [39] Qian J., L. Qian, N. Pu *et al.* An Intelligent Early Warning System for Harmful Algal Blooms: Harnessing the Power of Big Data and

- Deep Learning. *Environmental Science & Technology*, 2024, **58**(35): 15607-15618. DOI: 10.1021/acs.est.3c03906.
- [40] Yu P., R. Gao, D. Zhang *et al.* Predicting coastal algal blooms with environmental factors by machine learning methods. *Ecological Indicators*, 2021, **123**: 107334. DOI: 10.1016/j.ecolind.2020.107334.
- [41] Wang CL, Xue L, Zhang YS *et al.* Urban Ozone Driving Factors Based on Explainable Machine Learning. *Environmental Science*, 2025: 1-15. DOI: 10.13227/j.hjlx.202404320. [王超龙, 薛莲, 张宜升等. 基于可解释性机器学习的滨海城市臭氧驱动因素. 环境科学, 2025, 1-15.]
- [42] Miao WM, Huang TL. Control effect of water lifting aeration system on cyanobacteria bloom in summer under different operating conditions: a case study of a reservoir in southern China. *Journal of Environmental Engineering Technology*, 2025, **15**(03): 904-914. DOI: 10.12153/j.issn.1674-991X.20240206. [缪威铭, 黄廷林. 扬水曝气系统不同运行工况对夏季蓝藻水华的控制效果——以南方某水库为例. 环境工程技术学报, 2025, 15(3): 904-914.]
- [43] Zhang M, Yang Z, Shi XL. Expansion and drivers of cyanobacterial blooms in Lake Taihu. *Journal of Lake Sciences*, 2019, **31**(2): 336-344. DOI: 10.18307/2019.0203. [张民, 阳振, 史小丽. 太湖蓝藻水华的扩张与驱动因素. 湖泊科学, 2019, 31(2): 336-344.]
- [44] Wang Q., T. Wang, S. Zhao *et al.* Comprehensive meteorological factors analysis and lag correlation study for cyanobacterial blooms in shallow plateau lake. *Ecological Indicators*, 2023, **153**: 110394. DOI: 10.1016/j.ecolind.2023.110394.
- [45] Yuan J., Z. Cao, J. Ma *et al.* Influence of climate extremes on long-term changes in cyanobacterial blooms in a eutrophic and shallow lake. *Science of The Total Environment*, 2024, **939**: 173601. DOI: 10.1016/j.scitotenv.2024.173601.
- [46] Kim J.H., J.-K. Shin, H. Lee *et al.* Improving the performance of machine learning models for early warning of harmful algal blooms using an adaptive synthetic sampling method. *Water Research*, 2021, **207**: 117821. DOI: 10.1016/j.watres.2021.117821.
- [47] Recknagel F., P.T. Orr, M. Bartkow *et al.* Early warning of limit-exceeding concentrations of cyanobacteria and cyanotoxins in drinking water reservoirs by inferential modelling. *Harmful Algae*, 2017, **69**: 18-27. DOI: 10.1016/j.hal.2017.09.003.
- [48] Myer M.H., E. Urquhart, B.A. Schaeffer *et al.* Spatio-Temporal Modeling for Forecasting High-Risk Freshwater Cyanobacterial Harmful Algal Blooms in Florida. *Frontiers in Environmental Science*, 2020, **8**. DOI: 10.3389/fenvs.2020.581091.
- [49] de J. Magalhães A.A., L.D. da Luz, T.R. de Aguiar Junior. Environmental factors driving the dominance of the harmful bloom-forming cyanobacteria *Microcystis* and *Aphanocapsa* in a tropical water supply reservoir. *Water Environment Research*, 2019, **91**(11): 1466-1478. DOI: 10.1002/wer.1141.
- [50] Mu M., Y. Li, S. Bi *et al.* Prediction of algal bloom occurrence based on the naive Bayesian model considering satellite image pixel differences. *Ecological Indicators*, 2021, **124**: 107416. DOI: 10.1016/j.ecolind.2021.107416.