J. Lake Sci.(湖泊科学),2027,39(4):000-000
DOI: 10.18307/2027.0412
©2027 by *Journal of Lake Sciences*

湖泊藻类生长的时滞驱动特征解析：融合随机采样与物理约束的可解释机器学习*

王万宾[2,3]，余艳红[1,4**]，严长安[2,3]，刘　芳[2,3]
(1：南京师范大学地理科学学院,气候系统预测与变化应对全国重点实验室，南京 210023)
(2：云南省生态环境科学研究院,昆明 650034)
(3：云南省高原湖泊流域污染过程与管理重点实验室，昆明 650034)
(4：云南省生态环境监测中心,昆明 650034)

摘　要：驱动因子对湖泊藻类生长的驱动存在时滞特性，现有方法难以在高维交互因子的驱动条件下同时精准识别时滞区间并阐释其驱动特征。为此，本研究提出了一种融合随机采样与物理约束建模的可解释机器学习框架，旨在系统解析藻类生长的时滞驱动特征。该框架首先通过随机采样构建多时滞特征集，利用 XGBoost 模型与 SHAP 值分析，自适应地识别驱动因子影响藻类生长的最优时滞窗口。进而，在模型训练中引入现有物理约束项，并通过分析约束模型下的特征依赖图（PDP），定量揭示驱动因子的非线性作用特征。本研究以中国云南高原不同营养状态的两个湖泊（重度富营养化的滇池和轻度富营养化的洱海）为案例进行验证。结果表明：（1）该框架对叶绿素 α 浓度变化具有较好的模拟性能，验证集对滇池与洱海的 R^2 达到 0.88 以上，MAPE 低于 14.6%，构建的模型对滇池和洱海分别可实现提前 4 天（R^2=0.79, MAPE=13.3%）和 6 天（R^2=0.82, MAPE=16.8%）的 Chl.*a* 高精度预测。（2）高藻情期水体理化与气象因子的时滞普遍大于低藻情期，主要归因于水体理化因子对藻类生长的累积驱动特征及高藻情期对光照的遮蔽效应。（3）滇池营养盐对叶绿素 α 的驱动时滞快于洱海，体现富营养条件下营养物质的即时消耗特征。（4）重度富营养化的滇池，其藻类生长主要受到水体理化特征（pH、DO）、营养物质（TN）与可氧化有机物（COD_{Mn}）的共同驱动；而处于轻度富营养化状态的洱海，则主要受到水体理化（水温、pH、DO）的驱动。本研究构建的框架为从“识别时滞”到“阐释特征”的探索提供新的可行路径，有助于发展更具过程理解与可解释性的藻华预测模型，并为针对不同湖泊类型制定差异化、前瞻性的生态管理策略提供科学依据。
关键词：藻类生长；时滞效应；可解释机器学习；随机采样；物理约束建模；滇池；洱海

Deciphering time-Lag driving characteristics of algal blooms in lakes: An interpretable machine learning approach with random sampling and physical constraints*

Wang Wanbin [2,3], Yu Yanhong [1,4*], Yan Changan [2,3]&Liu Fang [2,3]
(1: *School of Geography, State Key Laboratory of Climate System Prediction and Risk Management, Nanjing Normal University, Nanjing* 210023, *P.R.China*)
(2: *Yunnan Academy of Ecological and Environmental Sciences, Kunming* 650034, *P.R.China*)
(3: *Yunnan Key Laboratory for Pollution Processes and Control of Plateau Lake-Watersheds, Kunming* 650034, *P.R.China)*
(4: *Yunnan Ecological and Environmental Monitoring Center, Kunming* 650034, *P.R.China*)

Abstract: Algal growth shows clear time-lag responses to driving factors. Current methods cannot accurately identify time-lag intervals or clarify the corresponding driving characteristics under high-dimensional interactions among multiple environmental factors. This study established a new interpretable machine learning framework. The framework combines random sampling and

*2026-03-03 收稿；2026-07-10 收修改稿；2026-07-17 录用。
国家自然科学基金项目（U2102207）、云南省科技厅重点研发计划（202503AT100008）、云南省基础研究计划项目（202501AT070436）、云南省生态环境科学研究院流域污染过程模拟新方法研究创新团队（2023-07CXTD）联合资助。
**通信作者，E-mail：342737000@qq.com。

physically constrained modeling to systematically analyze the time-lag driving characteristics of algal growth. It first builds a time-lag feature set through random sampling. It then uses XGBoost and SHAP analysis to adaptively determine the optimal time-lag window of each driving factor. Physical constraints are further introduced into model training. Partial dependence plots (PDP) based on the constrained model quantitatively reveal the nonlinear effects of driving factors. This study verified the framework using two plateau lakes with different trophic levels in Yunnan, China. These lakes are hypereutrophic Lake Dianchi and slightly eutrophic Lake Erhai. The results show four key findings. (1) The framework achieves high accuracy in predicting chlorophyll-a concentrations, with R^2 values higher than 0.88 and MAPE lower than 14.6%. The framework realizes reliable chlorophyll-a forecasting for Lake Dianchi at a 4-day lead time (R^2 = 0.79, MAPE = 13.3%) and for Lake Erhai at a 6-day lead time (R^2 = 0.82, MAPE = 16.8%). (2) Hydrochemical and meteorological factors show longer time lags during severe algal bloom periods than during low-bloom periods. This difference mainly results from the cumulative effects of hydrochemical factors on algal growth and from light attenuation induced by dense algal biomass. (3) Nutrient variations respond more rapidly to chlorophyll-a changes in Lake Dianchi than in Lake Erhai. This pattern indicates rapid nutrient consumption under hypereutrophic conditions. (4) Algal growth in hypereutrophic Lake Dianchi is jointly driven by hydrochemical factors (pH, DO), nutrients (TN), and oxidizable organic matter (COD_{Mn}). By comparison, algal growth in slightly eutrophic Lake Erhai is primarily controlled by hydrochemical factors (water temperature, pH, DO). The proposed framework provides a reliable approach for time-lag recognition and mechanistic interpretation. It supports the development of process-based and explainable algal bloom prediction models. It also offers scientific references for targeted and proactive ecological management of lakes with different trophic statuses.
Keywords: Algal growth; Time-lag effect; interpretable machine learning; random sampling; physically constrained modeling; Lake Dianchi; Lake Erhai

湖泊富营养化及其引发的有害藻华是全球性的水环境挑战。中国是全球湖泊数量最多的国家之一，其境内众多湖泊长期存在藻华（水华）现象[1]。过去数十年间，以控制外源营养盐（氮、磷）输入为核心的管理策略虽已取得一定成效，然而在一些湖泊中，藻华的发生频率、强度和持续时间并未得到根本性控制，反而在气候变暖背景下有加剧趋势[2-3]。这一现象无法仅用传统基于稳态假设和线性驱动的管理模型来解释。究其原因，其忽视了环境驱动因子对藻类生长影响的非线性和时滞性这两个关键特征[4-5]。

时滞效应是生物环境变化对藻类种群动态驱动的核心属性。气象因子（如气温、日照、风速）需通过改变水体热力学结构、光场和混合层深度来间接影响藻类生理；营养盐（如总磷、总氮）的效应则涉及从外源输入、沉积物释放、到生物吸收利用的复杂生物地球化学过程链[6-7]。例如，有研究表明气温对滇池藻华的驱动存在 25-30 天的滞后，对降水的驱动滞后约 10 天[8]。洱海的研究也指出，藻类种群对水温等因子存在显著的滞后特征[9]。这种时滞性的存在，使得基于同步数据的相关性分析或预测模型往往难以捕捉真实的驱动关系，导致预警滞后和管理措施失准[7]。

近年来，机器学习方法因其强大的非线性拟合能力，在藻华预测和驱动因子识别中展现出巨大潜力[5,10]。然而，主流机器学习模型（如 XGBoost）常被视为“黑箱”，其预测结果缺乏明确的物理解释，难以回答“为何驱动”及“滞后多久”等机理问题[10]。尽管 SHAP（SHapley Additive exPlanations）等可解释性工具能评估特征重要性，但它们通常基于同步数据，无法直接揭示时滞结构[3]。同时，纯数据驱动的模型可能学习到统计伪关联，而忽略了生态学的基本物理约束（如营养盐对藻类生长的生理驱动），导致模型外推能力差、机理洞察有限[7]。

为克服上述挑战，本研究提出一种融合随机采样与物理约束建模的可解释机器学习框架，旨在系统解析湖泊藻类生长的时滞驱动特征。该框架的创新包括以下两点：（1）基于随机采样的时滞区间识别。通过构建包含不同滞后时间序列特征的扩展数据集，并应用机器学习模型结合特征重要性 SHAP 分析，自适应地识别各驱动因子影响藻类生长的最优时滞窗口。（2）基于物理约束的驱动特征阐释。在模型训练中，引入基于藻类生长动力学的物理约束项，引导模型学习符合生理生态学原理关系，再通过分析约束模型下各驱动因子的特征依赖图（PDP），定量揭示驱动因子在特定时滞下对藻类生长的非线性影响模式。本研究以云南高原两个典型但营养状态差异显著的湖泊（重度富营养化的滇池和轻度富营养化的洱海）为案例，

通过应用本框架，旨在定量比较两湖藻类生长关键驱动因子及其时滞特性的差异，并评估模型的预测性能，以期为两湖差异化、精准化的藻华预测应用提供基于过程理解的科学依据。

1 研究区概况、数据来源与研究方法

1.1 研究区概况

本研究选取的两个代表性湖泊滇池和洱海（图 1），在营养状态、水文特征等方面存在显著差异[11-12]，为对比研究藻类生长的时滞驱动特征提供了理想对象。滇池位于云南省昆明市，是我国西南地区最大的淡水湖，也是中国第六大淡水湖，素有“高原明珠”之称。湖体面积约 300 km^2，平均水深约 4.7 m，属典型的浅水湖泊。自 20 世纪 80 年代末以来，滇池经历了严重的富营养化过程，是我国重点治理的三大富营养化湖泊之一[13]。洱海位于云南省大理白族自治州，是云南高原另一重要湖泊，湖面面积约 250 km^2，平均水深约 10.5 m，属高原断陷湖泊。与重度富营养化的滇池相比，洱海整体水质相对较好，处于轻度富营养化状态[14]。近年来，随着流域污染拦截、“三线”管控等治理措施的推进，洱海水质的季节波动性有所减弱，系统韧性增强，但夏秋季（汛期）因降雨带来的营养盐输入仍可能导致浮游植物生物量（以叶绿素 α 浓度为指示）升高。滇池与洱海分别代表了长期高营养盐负荷和营养盐相对受限但受气候变化敏感的两种典型高原湖泊生态系统，其藻类生长的主导驱动因子可能存在区别，为深入探究不同营养状态下藻类生长的时滞驱动特征提供了典型案例。

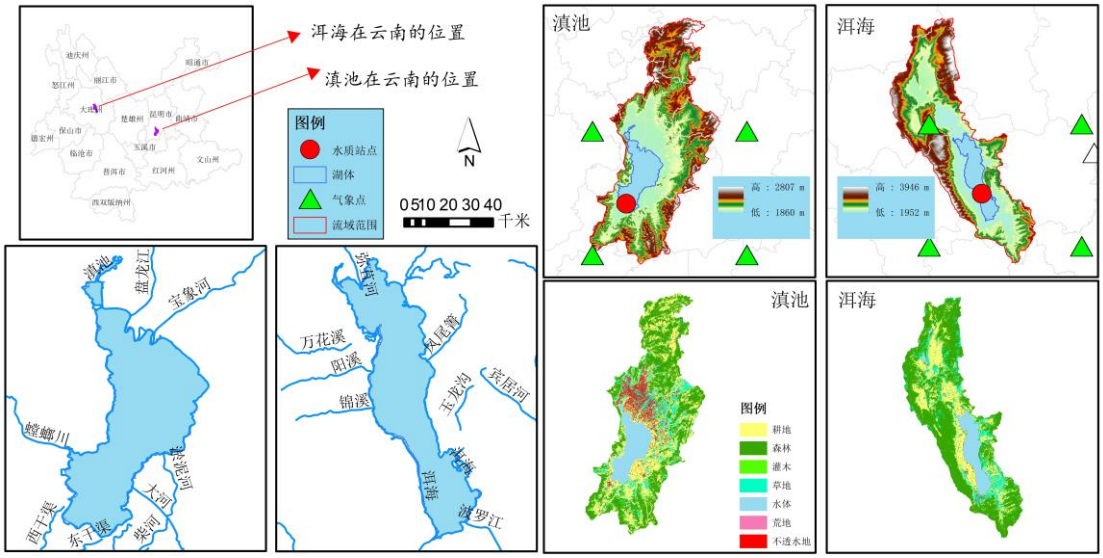


图 1 研究区域位置示意图

Figure 1 Location map of the study area

1.2 数据来源与预处理

以叶绿素 *a*（Chl.*a*）浓度表征藻类生物量。藻类生物量变化受多因子影响，主要包括气象因子、水体理化因子、营养盐以及可氧化有机物指标。其中，气象因子涵盖降雨量（PRE）、风速（WIND）、长波辐射（longwave radiation，L_rad）和短波辐射（shortwave radiation，S_rad）；水体理化因子包括溶解氧（dissolved oxygen，DO）、pH、水温（water temperature，Water_tem）和浊度（turbidity，TUR）；营养盐相关指标包括总氮（total nitrogen，TN）、氨氮（ammonia nitrogen，NH_3-N）、总磷（total phosphorus，TP）以及氮磷比（TN/TP ratio， NP_ratio）；可氧化有机物指标为高锰酸盐指数（permanganate index，COD_{Mn}）。藻类生物量及水环境指标来源于国家地表水水质自动监测实时数据发布系统（https://szzdjc.cnemc.cn/）。气象数据来源于欧洲中期天气预报中心（https://cds.climate.copernicus.eu/），

采用湖泊邻近 4 个网格点的平均气象数据。由于长波辐射与水温相关性较强（相关系数 ρ＞0.9），故舍弃长波辐射指标（见附图I、II）。所有数据的时间范围为 2021—2024 年，时间分辨率为逐日。各指标的缺失率均低于 5%，对缺失值采用线性插值进行填补。2021—2023 年期间的数据用于模型的构建及验证，2024 年的数据用于评价预测性能。根据 Chl.*a* 浓度的峰值变化特征，本研究划分了高藻情期（5—9 月）和低藻情期（10 月—次年 4 月），以便阐述不同藻期的时滞驱动差异性。为消除测量噪声与随机波动、提高数据质量，采用小波变换对数据进行去噪处理。小波去噪方法能够保留信号中的突变信息，因此极端事件（如暴雨、极端高温等）所对应的 Chl.*a* 高值会给予保留。本研究对比了去噪前后极端高值的时间序列，确认其幅值和时序位置基本保持不变（见附图III、IV）。具体流程如下：首先，利用 Daubechies 4（db4）小波基函数对原始数据进行 3 层分解；其次，采用通用阈值估计并结合软阈值函数对小波系数进行处理，以有效分离信号与噪声；最后，通过小波重构获得去噪后的数据序列。

1.3 研究方法

1.3.1 XGBoost 机器学习方法 XGBoost（eXtreme Gradient Boosting）是一种基于梯度提升框架的集成学习算法，通过逐步添加决策树来拟合前一轮预测的残差，并引入正则化项控制模型复杂度，具有高效、鲁棒且不易过拟合的优点[15]。本研究基于 Python 3.11 平台，采用 XGBoost 回归器（XGBRegressor）建立 12 个驱动因子与 Chl.*a* 浓度之间的非线性映射关系。模型中的 nestimators、max_depth、learning_rate、subsample 以及 colsample_bytree 共 5 个超参数通过贝叶斯优化进行选择（优化目标为 5 折交叉验证的负均方误差）。使用决定系数（R^2）、平均绝对百分比误差（MAPE）、均方根误差（RMSE）评估指标来评估模型的模拟和预测性能。

1.3.2 融合随机采样和 SHAP 分析的可解释机器学习方法 本研究构建循环采样与时滞分析集成方法，通过多重随机采样与 SHAP 可解释机器学习相结合，系统探究 12 个驱动因子在高藻情期与低藻情期对 Chl.*a* 的时滞影响效应。方法核心包括四个关键环节：（1）滞后特征设置。对 12 个驱动因子均设置最大时滞天数为 60d。（2）循环随机采样。在 0-60 天中随机选择每个因子的滞后期，分别通过 200、500、1000 次独立循环，构建多样化特征组合，评估不同循环采样次数对时滞期的影响。（3）XGBoost-SHAP 协同分析。数据集统一标准化后，采用贝叶斯超参优化、5 折交叉验证的 XGBoost 模型进行模拟。利用 SHAP 值量化每个滞后特征的重要性，从模型可解释性角度评估因子影响的稳健性。（4）时滞区间优化决策（图 2）。基于各驱动因子在 0-60 天内 SHAP 值的统计分布，提取单因子数据按时滞排序，利用 Savitzky-Golay 平滑技术进行处理，平滑后能更稳定地识别显著区间。显著区间核心判别阈值为中位数与标准差之和，选择平均 SHAP 值最高的区间作为最优时滞区间。最后，计算最优时滞区间的可信度。

SHAP 值是 Shapley 解释法的核心指标，能量化每个驱动因子对模型模拟结果的贡献度[16]。在每一轮随机时滞组合建模后，利用 shap.TreeExplainer 在对应数据集上计算各滞后特征的 SHAP 值，并以平均绝对 SHAP 值反映该时滞在当前组合下的边际预测贡献。对于 Chl.*a* 模拟模型$f(x)$和 12 个输入特征因子向量$x=(x_1,x_2,\ldots,x_M)$，特征j的 SHAP 值计算公式为：

$$\varnothing_j f(x)=\sum_{S\in\{1,2,\ldots,M\}\backslash\{j\}}\frac{|S|!(M-|S|-1)!}{M!}\left[f_x(S\cup\{j\})-f_x(S)\right] \tag{1}$$

式中：S是特征子集（不包含特征j），M是特征总数（12 个），$f_x(S)$是在特征子集S上的模型预测值。

Savitzky-Golay 法是基于局部最小二乘多项式拟合的平滑分析方法[17-18]。对于 0-60 天的 SHAP 系列x_i，在每个时滞点i处取窗口大小为 5，拟合一个 2 阶多项式（公式 2），通过最小二乘法确定系数，然后用多项式在中心点$t=0$处的值作为平滑结果。

$$p(t)=a_0+a_1t+a_2t^2 \tag{2}$$

本研究的可信度指标C由稳定性分类（基于变异系数CV）与区间长度分量加权构成[19]。计算公式如下：

$$\mu_I=\frac{1}{n_I}\sum_{i\in I}S_i \tag{3}$$

$$\sigma_I = \sqrt{\frac{1}{n_I}\sum_{i\in I}(S_i - \mu_I)^2} \tag{4}$$

$$CV = \begin{cases} \frac{\sigma_I}{\mu_I}, & \mu_I > 0 \\ 1, & \mu_I = 0 \end{cases} \tag{5}$$

$$C = \min[1, 0.7\max(0, 1 - CV) + 0.3\min\left(1, \frac{L_I}{0.2 * L_T}\right)] \tag{6}$$

式中，C为可信度，取值[0,1]，越高表示最优时滞区间越稳定，越可靠。当$C \geq 0.7$为高可信。CV 为变异系数，反映区间内 SHAP 值的相对离散程度。σ_I为S_i序列的标准差。μ_I为S_i序列的均值。I为检测出的最优滞后区间，包含n_I个滞后值。S_i为区间内第i个滞后的平均 SHAP 值。L_I为区间覆盖的滞后天数范围。L_T为全部滞后序列的长度(60d)。

1.3.3 融合物理约束和 PDP 分析的可解释机器学习方法　在最优时滞区间确定的基础之上，采用单调约束集成方法，将藻类生长规律与机器学习模型相结合，系统探究了 12 个环境因子对 Chl.*a* 的时滞驱动特征。方法设计包含三个关键环节：（1）基于前期分析结果，为每个环境因子预先定义最优时滞区间，随后在区间内进行随机采样 10 次，既保证了时滞选择的科学性，又考虑了不确定性。（2）物理规律约束的模型构建。在 XGBoost 模型中引入单调性约束。设定 NH_3-N、TP 和 TN 对藻类密度具有正向单调约束[20]，反映了这些因子促进藻类生长的基本生态原理（比如 Monod 动力学方程[21-22]）。（3）贝叶斯优化与 PDP 协同分析。采用贝叶斯优化自动搜索最佳模型超参数，同时结合 PDP 分析各驱动因子对 Chl.*a* 的非线性过程效应。（4）模型的预测性能分析。基于以上过程构建的最优模型，利用 2024 年独立数据集评估模型的预测性能。研究技术路线见图 2。

由于 PDP 可通过固定其他因子的值并逐渐改变关注驱动因子的值，显示单个因子和目标变量之间的关系[23-24]，从而揭示该因子和目标变量之间的非线性关系，因此特引入 PDP 实现 12 个驱动因子的驱动特征分析。对于特征x_s的部分依赖函数为：

$$f_s(x_s) = \frac{1}{n}\sum_{i=1}^{n} f(x_s, x_c^{(i)}) \tag{7}$$

式中, x_s是目标特征，x_c是其他特征，n是样本数量，f是 XGBoost 模型。

2 结果与分析

2.1 驱动因子的时滞特性

基于随机采样的时滞识别模块，量化了滇池与洱海 12 个驱动因子对 Chl.*a* 的时滞驱动特征（图 3）。经 500 与 1000 次循环验证，两湖时滞可信度变化仅 0.01 和 0.04，500 次循环足以稳定识别。水体理化因子（水温、pH、TUR）方面，两湖高藻情期时滞均长于低藻情期，推测为持续累积驱动所致；滇池低藻情期 DO 时滞尤其明显，可能与有机质分解耗氧滞后有关。营养盐（NH_3-N、TN、TP、N/P 比）中，滇池高藻情期多数时滞短于洱海，呈“即时消耗”特征，或与富营养下蓝藻不受氮磷限制有关[25-26]。气象因子上，PRE 和 S_rad 高藻情期时滞均大于低藻情期，辐射因遮蔽效应呈累积滞后，低藻期则直接促光合作用。滇池降雨时滞显著小于洱海，可能缘于城市地表污染物即时冲刷输入。风速对滇池高藻情期时滞短于低藻情期，可能与物理扰动快速改变藻类分布并促营养供给有关[27-28]；洱海则相反，高藻情期风速时滞更长，推测受营养盐与风速共同限制。有机物 COD_{Mn} 的时滞在两湖表现相反：滇池高藻情期大于低藻情期，洱海反之。总体而言，洱海对水环境和气象因子响应更超前，滇池对营养盐更超前。与 Wang 等[8]的气温、降水、风速时滞（25~30 天、<10 天、15~20 天）差异可能源于方法不同。渠姗等[9]在太湖的小时级时滞、Xie 等[7]的周尺度（7 周/4 周）差异，均表明湖泊特性、数据尺度及分析方法影响结果，未来需加强跨尺度对比。

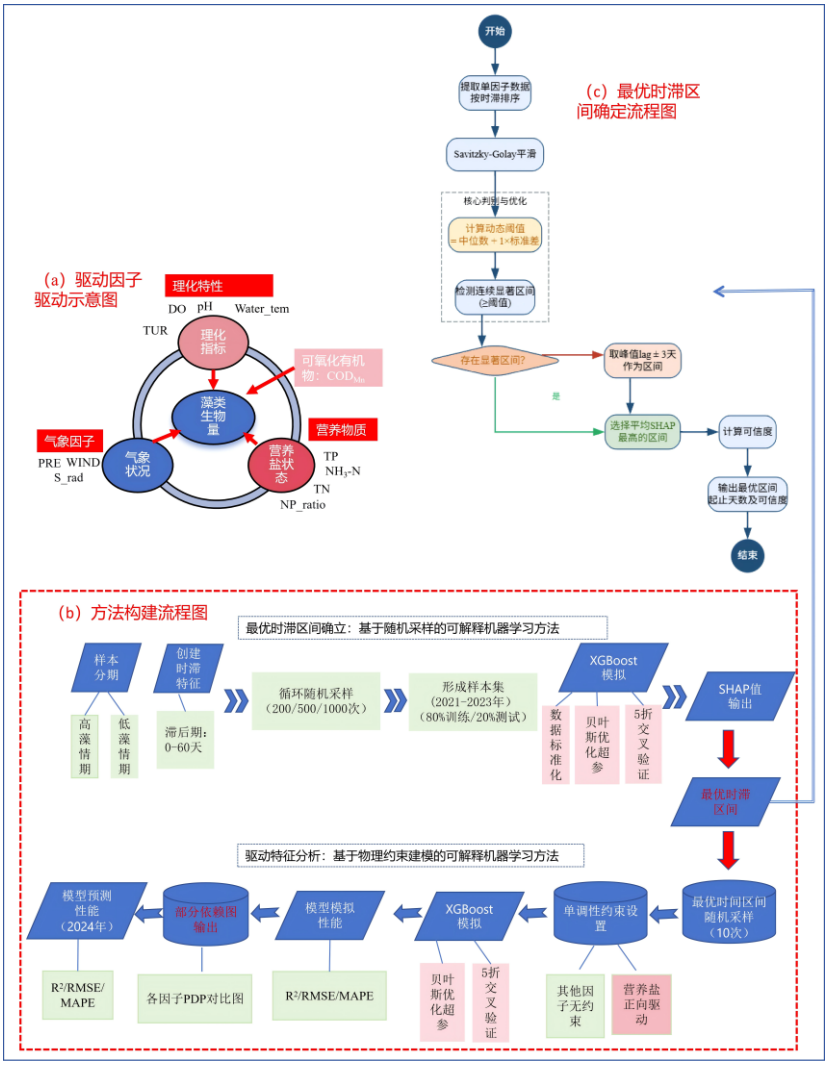


图 2　研究技术路线示意图

Fig.2 Schematic diagram of the research technical pathway

2.2 Chl.*a* 浓度模拟性能

在最优时滞区间内，利用 2021—2023 年期间的验证集随机模拟 10 次后滇池和洱海高藻情期和低藻情期验证集 R^2 均大于 0.88，MAPE 均控制在 14.6%以内，最优时滞区间内模型具有稳定并且较高的拟合精度（图 4）。本研究基于最优时滞区间及 2021—2023 年构建的模型，针对 2024 年，滇池能实现 Chl.*a* 提前 4 天模拟预测，R^2=0.79，MAPE=13.3%。针对洱海能实现 Chl.*a* 提前 6 天模拟预测，R^2=0.82，MAPE=16.8%（图 5）。对于富营养化湖泊巢湖，Zhang 等[29]研究构建了融合时间累积效应的机器学习框架，能够实现未来“1 天、3 天、7 天”的蓝藻水华空间分布预测，其中随机森林模型在 1 天预测中精度最高（R^2=0.81），空间预测像素精度达 84%，但预测时长延长至 7 天时精度（R^2<0.6）会显著下降。而李未等[30]利用千岛湖国控监测断面小金山 2016—2019 年 Chl.*a* 浓度的高频时间序列数据，构建了基于 NARX 神经网络的藻类预测模型，用于预测未来 0.5~7 天 Chl.*a* 浓度的变化，R^2 保持在 0.8~0.9 之间。Villanueva 等[31]以微囊藻毒素浓度为目标变量，将各种生物学、化学、气候和土地利用参数作为输入变量，建立一个提前一周预测有害藻华事件的分类模型，发现不同机器学习方法预测准确率均大于 0.80。Ai 等[32]基于分类和回归模型，

实现了伊利湖短期（10 天尺度）藻华暴发的预测，并在测试集中实现了高达 89.6%的准确率。这些研究表明，藻类模拟预测的时长跨度可从数天（短期预警）延伸至数周（中期预测），其精度高度依赖于模型方法、是否考虑环境因子的时滞/累积效应以及数据的时空覆盖度。本研究采用的基于驱动因子最优时滞区间的 Chl.*a* 预测方法，不仅在认知时滞特征上具有明显优势，而且模拟精度较高。

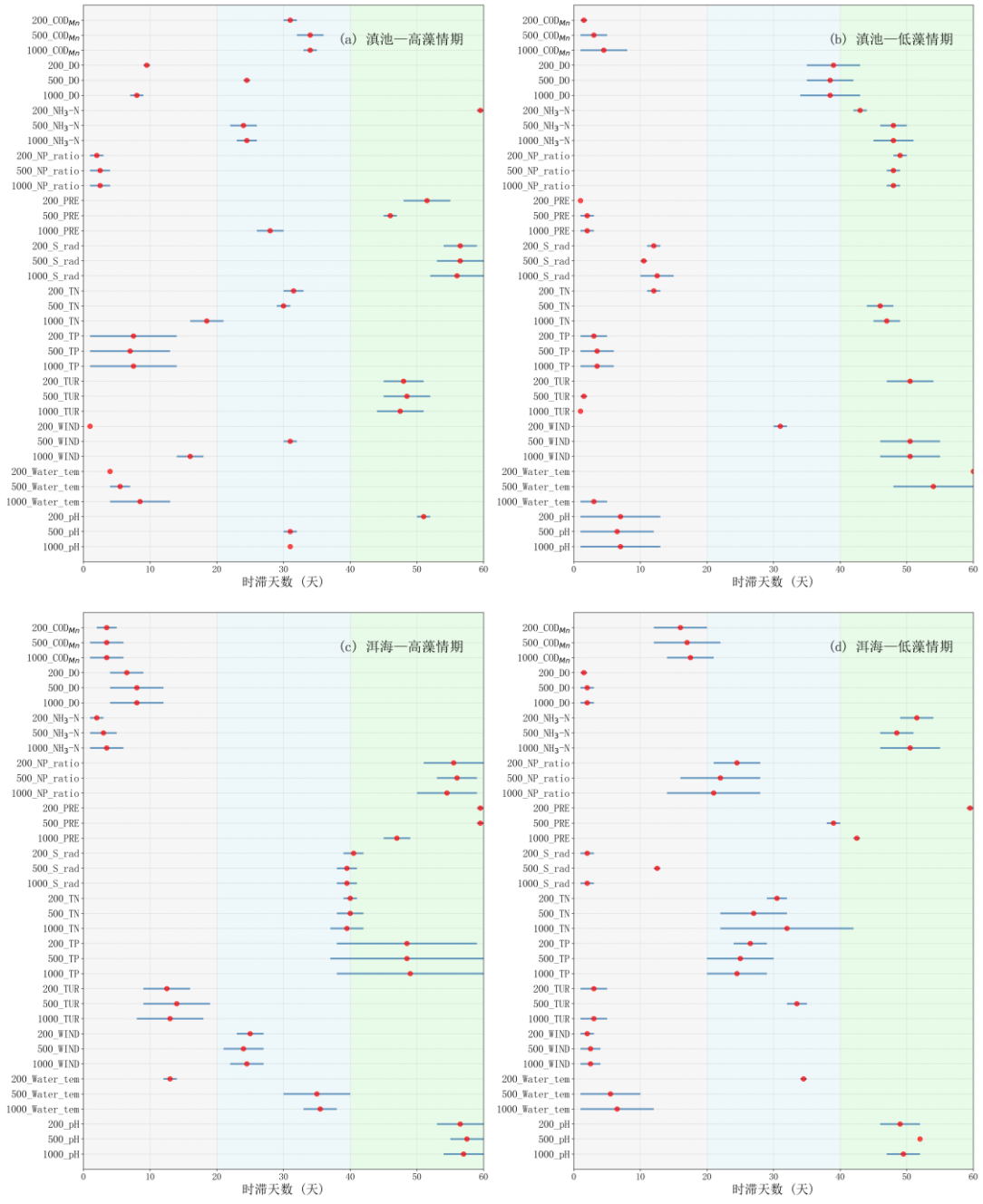


图 3 不同循环采样（200/500/1000 次）下驱动因子对高藻情期与低藻情期 Chl.*a* 的最优时滞性区间

Fig.3 Optimal time-lag intervals of driving factors for Chl.*a* during high-algae and low-algae periods under different bootstrap sampling times (200/500/1000)

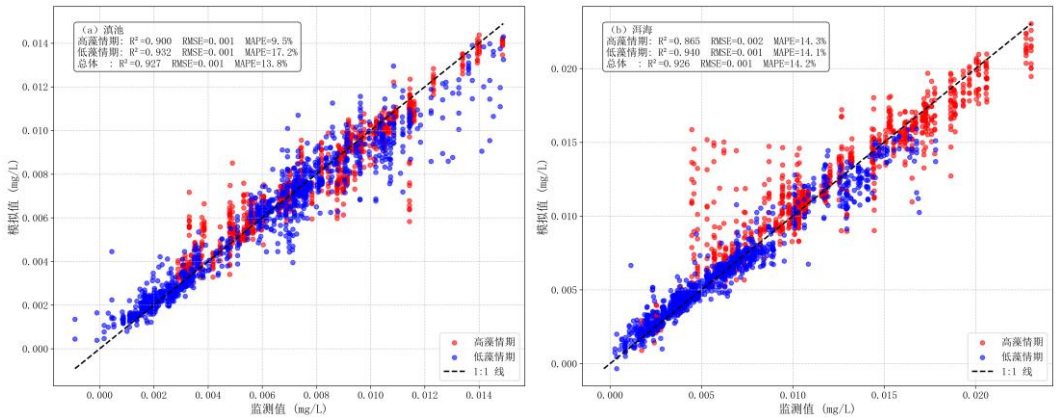


图 4 基于驱动因子时滞特征的 Chl.*a* 浓度模拟性能分析（2021—2023 年）

Fig.4 Simulation performance analysis of Chl.*a* concentration based on the time-lag characteristics of driving factors (2021-2023)

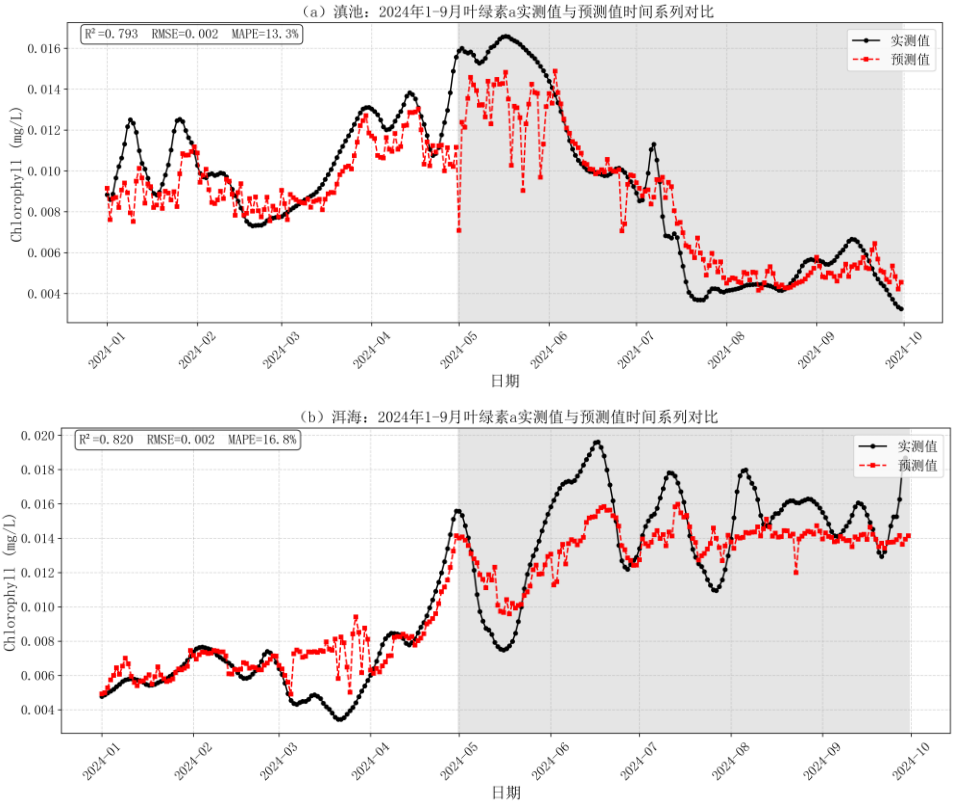


图 5 基于驱动因子时滞特征的 Chl.*a* 浓度预测性能分析（2024 年）

Fig.5 Analysis of predictive performance for Chl.*a* concentration based on time-lag characteristics of driving factors, 2024

2.3 驱动因子的驱动特征

基于物理约束与 PDP 分析，识别出滇池与洱海 Chl.*a* 的驱动因子存在明显差异（图 6 和图 7）。滇池主要受 pH、DO、TN、NP_ratio 和 COD_{Mn} 影响，其中 N/P 比高于 20 时抑制藻类，可能是由于过量 N 促

使藻类合成含氮毒素（如微囊藻毒素），导致生长代价上升[33]。而 COD_{Mn} 在长期高营养盐背景下成为蓝藻增殖的关键辅助因子，可能与其化学组分中易降解有机部分作为碳源被吸收有关，这不同于传统认为磷或氮为绝对限制因子的观点[34-35]。洱海则以水温、pH、DO、TUR 为主导，水温超过 20 °C 后抑制藻类，印证了热分层和代谢速率调控的重要性[2-3,7]。在共性层面上，高 pH 和低 DO 通常协同促进藻类生长[5]。其中，高 pH 通过改变无机碳（CO_2/HCO_3^-）平衡有利于蓝藻类群的碳吸收；而低 DO 则可能通过削弱鱼类的捕食压力及底栖动物的扰动作用，间接缓解浮游植物的摄食限制，从而为其藻类种群增长创造有利条件。滇池极高 pH（>9.0）可能产生毒性抑制；DO 高于 8 mg/L 时 Chl.*a* 下降趋缓，可作为指示因子，可能与藻类呼吸等代谢耗氧对 DO 变化的响应减弱有关，这与部分研究发现的复杂关系并不矛盾[36]。光照是驱动两湖藻类增殖的必要环境条件，与稳态转换理论的基本预期一致[34,37]。而降水在两湖中驱动效应不显著，这一结果不同于部分报道[3,8]，可能归因于两湖较大的环境容量掩盖了流域径流输入的营养盐波动。

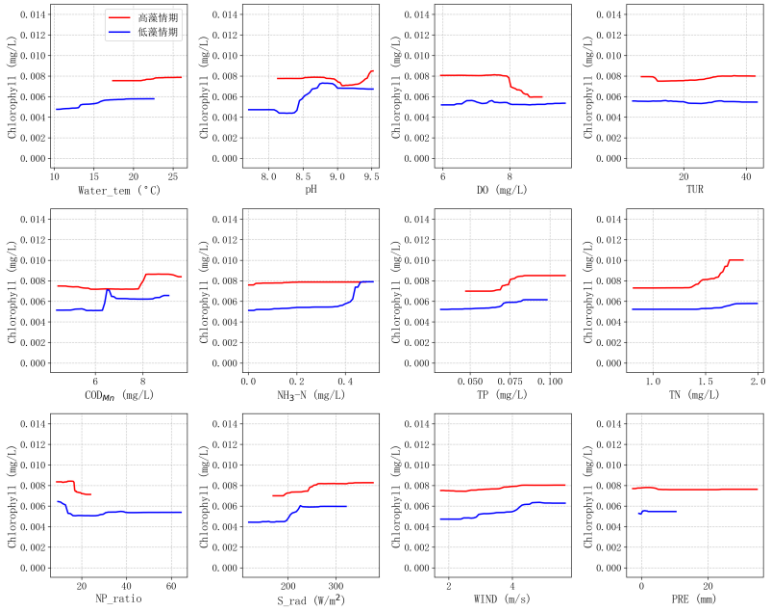


图 6 滇池驱动因子驱动 Chl.*a* 浓度变化的非线性特征

Fig.6 Nonlinear characteristics of driving factors on Chl.*a* concentration changes in Lake Dianchi

3 讨论

准确刻画环境因子对 Chl.*a* 浓度的时滞驱动，是构建高精度藻华预测模型和理解湖泊生态系统非线性动力学的基础环节。本研究提出并应用了一种基于随机采样与 SHAP 分析的可解释机器学习方法，旨在实现时滞窗口的自适应、数据驱动式识别。相较于传统方法，该策略不依赖先验假设或固定窗口，能够从数据中自动筛选出最具预测力的时滞组合，尤其适用于多因子协同、时滞交织的复杂场景；同时，随机采样生成的丰富特征组合使模型可隐性考察不同因子在不同滞后时间下的交互作用，而 SHAP 分析则提供了关键驱动因子的可视化解释，兼顾了预测能力与可解释性。

将上述方法应用于滇池与洱海，结果显示两湖在营养状态和驱动结构上呈现显著分化。重度富营养化的滇池，其 Chl.*a* 动态受水体理化特征（pH、DO）、营养盐（TN）与可氧化有机物（COD_{Mn}）的共同驱动，多因子之间形成协同放大效应；这一模式与太湖梅梁湾等富营养化湖区的观测一致[1,9]，表明在营养盐过饱和条件下，水温、pH 等理化因子通过调控藻类代谢和营养盐形态，与营养盐共同构成限制性条件，甚至使气候因子的贡献日益凸显。相比之下，轻度富营养化的洱海则主要由水温、pH 和 DO 等理化因子主导，营养盐的直接影响较弱，类似于呼伦湖等贫营养或寒旱区湖泊的情形[38]，反映出水体物理化学环境的稳定性对维持低营养状态具有关键作用。上述对比提示，针对滇池的治理需在持续削减外源营养盐的同

时，强化生态补水、人工增氧和清淤等内源及理化环境调控措施，以打断营养盐与不利水体条件之间的正反馈；而洱海则应将重心置于预防营养盐进一步升高，并积极应对气候变化对水温分层和物理生境的潜在冲击，从而提升湖泊生态系统的韧性与稳定性。将上述基于最优时滞区间的方法应用于滇池与洱海的 Chl.*a* 预测，2021—2023 年验证集多次随机模拟，两湖高、低藻情期 R^2 均大于 0.88、MAPE 低于 14.6%，表明模型在拟合稳定性上表现优异。2024 年独立预测中，实现滇池提前 4 天和洱海提前 6 天的高精度预测，反映时滞区间优化对延长有效预警窗口的关键作用。

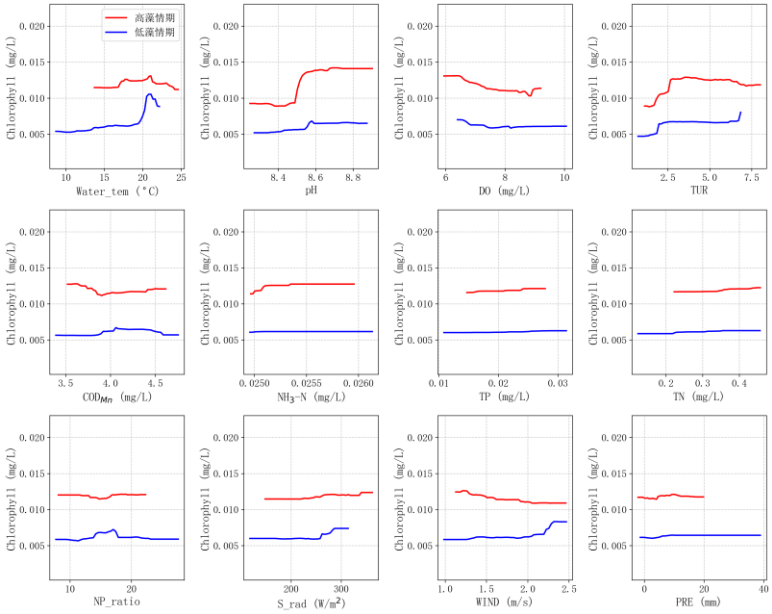


图 7 洱海驱动因子驱动 Chl.*a* 浓度变化的非线性特征

Fig.7 Nonlinear characteristics of driving factors on Chl.*a* concentration changes in Lake Erhai

尽管本方法在时滞自适应识别、特征解释和预测性能方面展现出优势，但其应用仍存在一些亟待深化的局限。首先，本研究主要依赖野外高频监测数据进行统计推断，所识别的时滞关系（如滇池 Chl.*a* 对高锰酸盐指数的滞后响应）实质上反映的是统计预测意义上的强关联，其背后可能是藻类对溶解性有机碳的直接利用、微生物矿化过程或碳氮磷代谢耦合等多种并行微观机制共同作用的结果，仅凭数据驱动方法难以推断因果关系。因此，未来亟需将本方法与室内控制实验和机理模型耦合，通过受控条件下的扰动实验验证特定滞后路径的因果链，从而弥补纯数据驱动方法在因果推断上的固有短板。其次，目前的方法验证主要基于滇池和洱海两个案例，尚缺乏与互相关分析[4,7,32]、格兰杰因果检验、交叉小波变换[9]等主流时滞识别方法在同一数据集上的系统定量比较，也缺少对适用场景（如不同时间分辨率、数据噪声水平、因子共线性程度）的边界条件评估。后续研究应拓展至更多类型、更广地域的湖泊水库，涵盖不同水文气候带和人类活动强度，以检验方法的普适性和稳健性。总之，本研究为时滞驱动的自适应识别提供了一条可解释机器学习的新路径，但其生态物理机制的落地仍需多学科证据的交叉支撑，这也将是推动数据驱动方法从“预测”走向“理解”的关键一步。

4 结论

本研究提出一种基于随机采样与物理约束建模的可解释机器学习方法框架，以滇池和洱海为案例，在驱动因子对 Chl.*a* 浓度变化上具有较好的模拟性能(R^2>0.88、MAPE 控制在 14.6%以内)。高藻情期水体理化与气象因子时滞普遍大于低藻情期，缘于累积驱动特征与藻类遮蔽效应。滇池营养盐时滞快于洱海，体现营养物质的即时消耗。重度富营养化的滇池，其藻类生长主要受到水体理化特征（pH、DO）、营养物质（TN）与可氧化有机物（COD_{Mn}）的共同驱动。而处于轻度富营养化状态的洱海，则主要受到

水体理化特征（如水温、pH、DO）的驱动。基于最优时滞模型，滇池和洱海分别可实现提前 4d（R^2=0.79, MAPE=13.3%）和 6d（R^2=0.82, MAPE=16.8%）的 Chl.*a* 高精度预测，验证了方法的有效性。

5 附录

附录Ⅰ~Ⅳ见电子版（DOI 10.18307/2027.0412）。

6 参考文献

[1] Cao J, Chu ZS, Jia HF *et al*. Succession of algae population and blooms with their driving factors in China's lakes. *Journal of Environmental Sciences*, 2026, **160**: 13-26. DOI: 10.1016/j.jes.2025.05.012.

[2] Chen QY, Wang SR, Ni ZK *et al*. No-linear dynamics of lake ecosystem in responding to changes of nutrient regimes and climate factors: Case study on Dianchi and Erhai lakes, China. *Science of the Total Environment*, 2021, **781**: 146761. DOI: 10.1016/j.scitotenv.2021.146761.

[3] Ma YM, Wang JF, Huo SL *et al*. Explainable machine learning reveals climate warming increases risk of algal blooms in lakes and reservoirs. *Water Research*, 2025, **287**: 124460. DOI: 10.1016/j.watres.2025.124460.

[4] Wang JK, He MQ, Pan M *et al*. Deciphering nonlinear effects of climate and eutrophication synergies on *Microcystis* blooms using a causal inference framework. *Harmful Algae*, 2025, **148**: 102911. DOI: 10.1016/j.hal.2025.102911.

[5] Cui H, Tao YW, Li J *et al*. Predicting and analyzing the algal population dynamics of a grass-type lake with explainable machine learning. *Journal of Environmental Management*, 2024, **354**: 120394. DOI: 10.1016/j.jenvman.2024.120394.

[6] Liu F, Zhang HG, Wang YB *et al*. Hysteresis analysis reveals how phytoplankton assemblage shifts with the nutrient dynamics during and between precipitation patterns. *Water Research*, 2024, **251**: 121099. DOI: 10.1016/j.watres.2023.121099.

[7] Xie YX, Chen SR, Zhou F *et al*. Development of a hybrid algal population prediction (HAPP) model by algae growth potential estimation and time series regression and its application in one reservoir in China. *Water Research*, 2025, **287**(Pt B): 124419. DOI: 10.1016/j.watres.2025.124419.

[8] Wang Q, Sun L, Zhu Y *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.

[9] 渠姗. 太湖叶绿素对气象因子时滞性响应及模拟研究[学位论文]. 南京：南京信息工程大学, 2024.

[10] Recknagel F, Shan K, Adrian R *et al*. Early warning of harmful cyanobacteria blooms based on high frequency *in situ* monitoring and intelligible machine learning modelling: The case study of Lake Müggelsee (Germany). *Water Research*, 2025, **287**: 124514. DOI: 10.1016/j.watres.2025.124514.

[11] Zhang HS, Jiao YM, Chen F *et al*. Multi-scale impacts of human activity intensity on water quality in nine plateau lake basins in Yunnan Province. *J Lake Sci*, 2024, **36**(2): 430-442. DOI: 10.18307/2024.0221. [张洪森, 角媛梅, 陈凡等. 云南九大高原湖泊流域人类活动强度对水质的多尺度影响. 湖泊科学, 2024, **36**(2): 430-442.]

[12] Ni ZK, Wang SR, Jin XC *et al*. Study on the evolution and characteristics of eutrophication in the typical lakes on Yunnan-Guizhou Plateau. *Acta Scientiae Circumstantiae*, 2011, **31**(12): 2681-2689. [倪兆奎, 王圣瑞, 金相灿等. 云贵高原典型湖泊富营养化演变过程及特征研究. 环境科学学报, 2011, **31**(12): 2681-2689.]

[13] Lou B, Duan Y, Huang J *et al.* Comprehensive diagnosis on eutrophication in Dianchi Lake. *Yangtze River*, 2025, **56** (12): 60-69. DOI:10.16232/j.cnki.1001-4179.2025.12.006.

[14] Hua ZH, Li R, Yang Z *et al*. The spatiotemporal variation and potential causes for the nutrient status of water in Lake Erhai from 2017 to 2022. *J Lake Sci*, 2024, **36**(6): 1639-1649. DOI: 10.18307/2024.0611. [华兆晖, 李锐, 杨智等. 2017—2022 年洱海水体营养状态的时空变化趋势及其成因分析. 湖泊科学, 2024, **36**(6): 1639-1649.]

[15] Sagi O, Rokach L. Approximating XGBoost with an interpretable decision tree. *Information Sciences*, 2021, **572**: 522-542. DOI: 10.1016/j.ins.2021.05.055.

[16] Qian ST, Qiao XY, Zhang WM *et al*. Machine learning-based prediction for settling velocity of microplastics with various shapes. *Water Research*, 2024, **249**: 121001. DOI: 10.1016/j.watres.2023.121001.

[17] Niedźwiecki MJ, Ciołek M, Gańcza A *et al*. Application of regularized Savitzky–Golay filters to identification of time-varying systems. *Automatica*, 2021, **133**: 109865. DOI: 10.1016/j.automatica.2021.109865.

[18] Hasan K, Meraj ST, Othman MM *et al*. Savitzky–golay filter-based PLL: Modeling and performance validation. *IEEE Transactions on Instrumentation and Measurement*, 2022, **71**: 2004306. DOI: 10.1109/TIM.2022.3196946.

[19] Zhang FM, Hua WJ, Li YR. Stability analysis of several comprehensive evaluation methods. *Journal of Systems Science and Mathematical Sciences*, 2019, **39**(4): 595-610. [张发明, 华文举, 李玉茹. 几种综合评价方法的稳定性分析. 系统科学与数学, 2019, **39**(4): 595-610.]

[20] Zhang S, Arhonditsis GB, Ji YL *et al*. Climate change promotes harmful algal blooms in China's lakes and reservoirs despite significant nutrient control efforts. *Water Research*, 2025, **277**: 123307. DOI: 10.1016/j.watres.2025.123307.

[21] 黄悦. 水中有机物和营养盐对藻类繁殖的耦合作用研究[学位论文]. 西安：西安建筑科技大学, 2021.

[22] Jiao JD, Zheng JF, Li FK *et al*. Analysis for growth characteristics and phosphorus nutrient dynamics of algae in mesohalinous water. *Environmental Pollution & Control*, 2017, **39** (7): 726-729, 734. DOI:10.15985/j.cnki.1001-3865.2017.07.007. [焦继东, 郑剑锋, 李付宽等. 混盐条件下藻类生长特性及磷营养盐动力学分析. 环境污染与防治, 2017, **39** (7): 726-729, 734.]

[23] Guo WZ, Liu J, Song YJ *et al*. Seismic fragility analysis of horizontally curved bridge portfolios using a machine-learning-based partial dependence plot approach. *Structure and Infrastructure Engineering*, 2025: 1-18. DOI: 10.1080/15732479.2025.2571603.

[24] Yosri AM, Sakthieswaran N, Ray S *et al*. Prediction of mechanical strength and performance of alkali-activated binders concrete using optimized machine learning with partial dependence plot. *Structures*, 2026, **84**: 110896. DOI: 10.1016/j.istruc.2025.110896.

[25] Wang Y, Zhao D, Woolway RI *et al*. Global elevation of algal bloom frequency in large lakes over the past two decades. *National Science Review*, 2025, **12**(3): nwaf011. DOI: 10.1093/nsr/nwaf011.

[26] Huang JC, Zhang YJ, Arhonditsis GB *et al*. The magnitude and drivers of harmful algal blooms in China's lakes and reservoirs: A national-scale characterization. *Water Research*, 2020, **181**: 115902. DOI: 10.1016/j.watres.2020.115902.

[27] 李沅睿. 基于多源遥感的滇池藻类水华物候与强度变化及其驱动因子研究[学位论文]. 昆明：云南大学, 2023.

[28] Li YR, Tao J, Zhang YL *et al*. Urbanization shifts long-term phenology and severity of phytoplankton blooms in an urban lake through different pathways. *Global Change Biology*, 2023, **29**(17): 4983-4999. DOI: 10.1111/gcb.16828.

[29] Zhang KL, Qi TC, Shen M *et al*. Time cumulative effects and machine learning–driven forecasting of cyanobacterial blooms in data-sparse eutrophic lakes. *Journal of Hydrology*, 2026, **668**: 134966. DOI: 10.1016/j.jhydrol.2026.134966.

[30] Li W, Zhu MY, Wang YC *et al*. Short-term forecasting model for algae based on NARX neural network in Qiandaohu Reservoir. *Research of Environmental Sciences*, 2022, **35**(4): 918-925. [李未, 朱梦圆, 王裕成等. 基于 NARX 神经网络的千岛湖藻类短期预测模型构建. 环境科学研究, 2022, **35**(4): 918-925.]

[31] Villanueva P, Yang J, Radmer L *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.

[32] Ai HP, Zhang K, Sun JC *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.

[33] Van de Waal DB, Smith VH, Declerck SAJ *et al*. Stoichiometric regulation of phytoplankton toxins. *Ecology Letters*, 2014, **17**(6): 736-742. DOI: 10.1111/ele.12280.

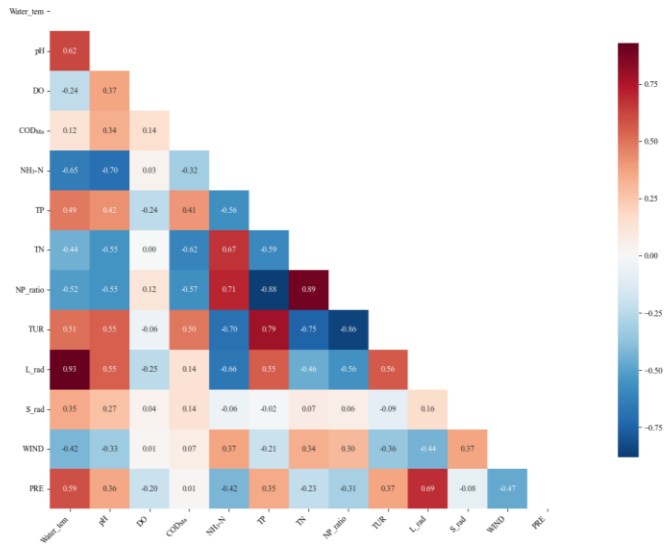
[34] Shi FN, Yang H, Liu XW *et al*. Spatial-temporal distribution of phytoplankton in Waihai of Dianchi lake and its nonlinear correlation analysis with environmental factors. *Acta Hydrobiologica Sinica*, 2022, **46**(7): 1070-1079. DOI: 10.7541/2022.2021.0159. [施凤宁, 阳辉, 刘湘伟等. 滇池外海藻类时空分布及其与环境因素间非线性相关分析. 水生生物学报, 2022, **46**(7): 1070-1079.]

[35] Scott JT, Taylor JM, Andersen IM *et al*. Reducing nitrogen inputs can rapidly reverse eutrophication in shallow phosphorus-rich lakes. *Environmental Science & Technology*, 2025, **59**(42): 22576-22586. DOI: 10.1021/acs.est.5c05392.

[36] Huang SL, Zang CJ, Du SL *et al*. Study on the relationships among pH, dissolved oxygen and chlorophyll a II: Non-aquaculture water. *Techniques and Equipment for Environmental Pollution Control*, 2011, **5**(8): 1681-1688. [黄岁樑, 臧常娟, 杜胜蓝等. pH、溶解氧、叶绿素 *a* 之间相关性研究II: 非养殖水体. 环境工程学报, 2011, **5**(8): 1681-1688.]

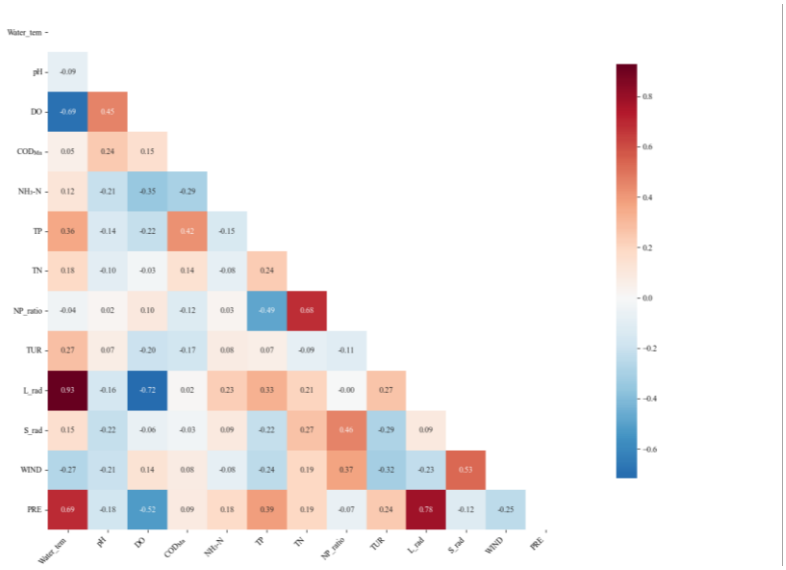
[37] Yang RB, Fan X, Zhao L *et al*. Identification of major environmental factors driving phytoplankton community succession before and after the regime shift of Erhai Lake, China. *Ecological Indicators*, 2023, **146**: 109875. DOI: 10.1016/j.ecolind.2023.109875.

[38] Guo ZY, Li CY, Shi XH *et al*. Spatial and temporal distribution characteristics of chlorophyll a content and its influencing factor analysis in Hulun Lake of cold and dry areas. *Ecology and Environmental Sciences*, 2019, **28**(7): 1434-1442. [郭子扬, 李畅游, 史小红等. 寒旱区呼伦湖水体叶绿素 *a* 含量的时空分布特征及其影响因子分析. 生态环境学报, 2019, **28**(7): 1434-1442.]



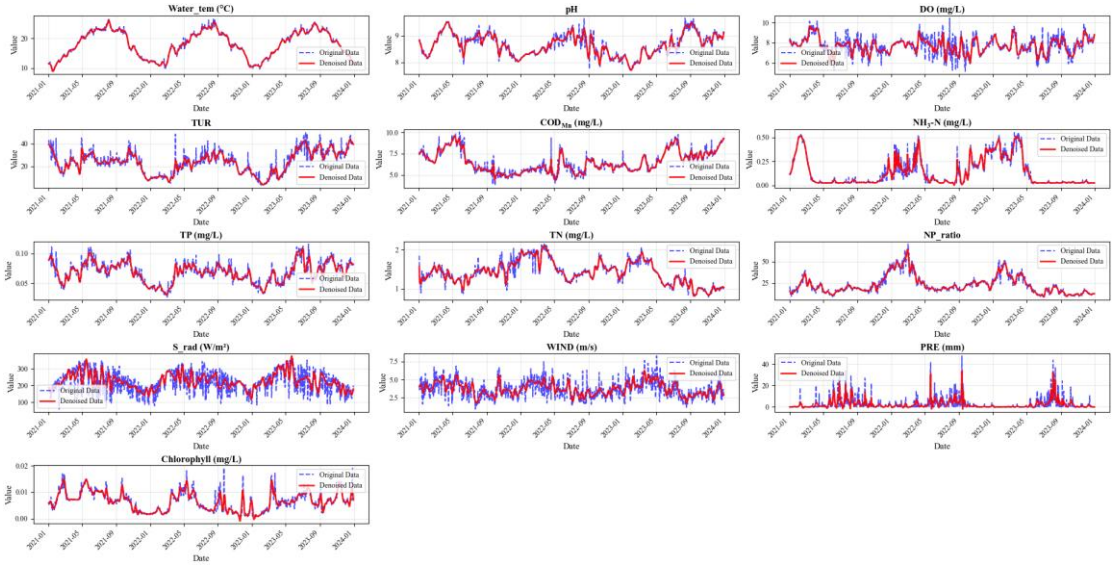
附图 I 滇池驱动因子之间的相关系数矩阵图

Attachted Fig.I Correlation coefficient matrix plot of driving factors in Dianchi Lake



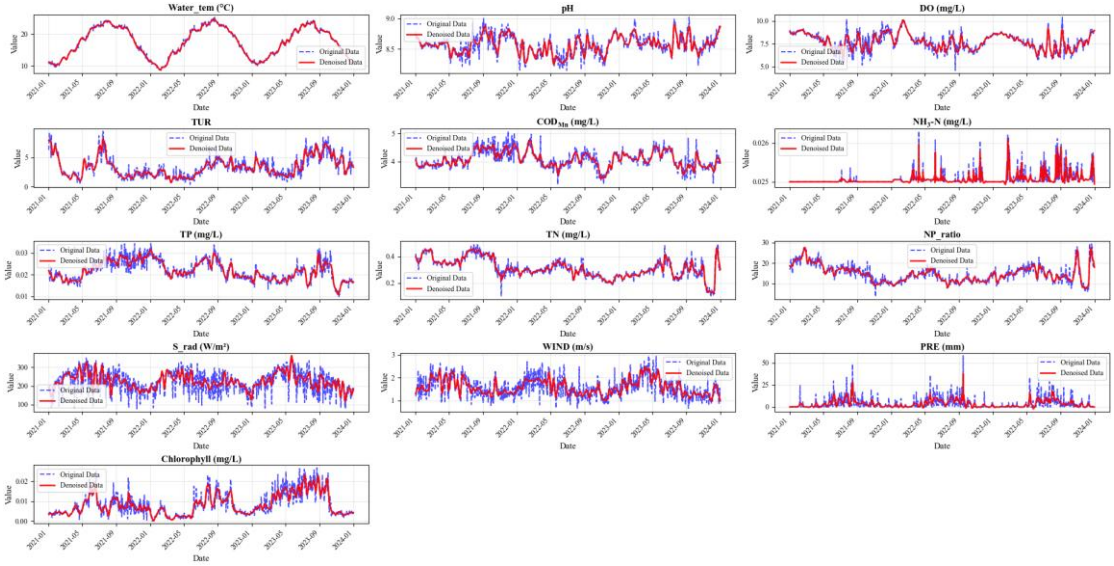
附图 II 洱海驱动因子之间的相关系数矩阵图

Attachted Fig. II Correlation coefficient matrix plot of driving factors in Erhai Lake



附图 III　2021-2023 年滇池 Chl.*a* 及各驱动因子小波去噪前后对比图

Attachted Fig. III Comparison of Chl.*a* chlorophyll and various driving factors in Dianchi Lake before and after wavelet denoising (2021–2023)



附图 IV　2021—2023 年洱海 Chl.*a* 及各驱动因子小波去噪前后对比图

Attachted Fig. IV Comparison of Chl.*a* chlorophyll and various driving factors in Erhai Lake before and after wavelet denoising (2021–2023)