

彭涛, 张晟, 祁少俊, 等. 基于 GCN 和 TPA 混合模型的青草沙水库叶绿素 a 浓度预测方法[J]. 净水技术, 2024, 43(9): 63-72.

PENG T, ZHANG S, QI S J, et al. Prediction method of Chlorophyll-a concentration of Qingcaosha Reservoir based on the hybrid model of GCN and TPA[J]. Water Purification Technology, 2024, 43(9): 63-72.

## 基于 GCN 和 TPA 混合模型的青草沙水库叶绿素 a 浓度预测方法

彭涛<sup>1</sup>, 张晟<sup>1</sup>, 祁少俊<sup>2</sup>, 张海平<sup>1,\*</sup>

(1. 同济大学环境科学与工程学院, 上海 200092; 2. 自然资源部东海海域海岛中心, 上海 200135)

**摘要** 青草沙水库是上海重要饮用水水源地, 面临较大藻华暴发风险, 实现对水库叶绿素 a (Chl-a) 浓度的准确预测, 对于饮用水安全保障至关重要。文章以青草沙水库为对象, 提出了一种图卷积时间模式注意力网络混合模型 (GC-TPA), 首先利用时间模式注意力 (TPA) 机制学习水质数据的时间依赖性, 其次使用图卷积网络 (GCN) 学习不同水质参数之间的关系。另外, 为进一步提高模型的预测精度, 引入完全自适应噪声集合经验模态分解 (CEEMDAN) 以降低模型的滞后性, 同时使用多层感知机 (MLP) 学习 Chl-a 浓度的突变。结果表明: (1) 引入的 GCN 模块显著增强了 TPA 对 Chl-a 的预测能力, 结合 CEEMDAN 和 MLP 的帮助, 模型性能进一步提升, 以纳什效率系数作为评价指标, 混合模型的 24 h 预测精度较单独 TPA 提升了 56.5%; (2) 与单独 TPA 和长短期记忆网络 (LSTM) 的对比试验表明, 在更长的预测周期 (48 h) 上, GC-TPA 虽然精度下降, 但仍表现最好, 48 h 预测平均绝对误差和均方误差比 LSTM 低 25.5% 和 24.0%, 比 TPA 低 4.92% 和 8.40%; (3) GCN 模块与 MLP 模块在结果预测中发挥了不同的作用, GCN 模块增强了 TPA 的特征学习能力, 提高了模型对 Chl-a 浓度变化趋势的预测精度, 而 MLP 则对 Chl-a 的突变较为敏感。研究所提出的 GC-TPA 混合模型在青草沙水库 Chl-a 浓度短期预测中表现良好, 可为水库水质管理提供支撑。

**关键词** 富营养化 叶绿素 a (Chl-a) 图卷积网络 (GPN) 时间模式注意力 (TPA) 多层感知机

**中图分类号:** TU991 **文献标识码:** A **文章编号:** 1009-0177(2024)09-0063-10

**DOI:** 10.15890/j.cnki.jsjs.2024.09.008

## Prediction Method of Chlorophyll-a Concentration of Qingcaosha Reservoir Based on the Hybrid Model of GCN and TPA

PENG Tao<sup>1</sup>, ZHANG Sheng<sup>1</sup>, QI Shaojun<sup>2</sup>, ZHANG Haiping<sup>1,\*</sup>

(1. College of Environmental Science and Engineering, Tongji University, Shanghai 200092, China;

2. East China Sea Area and Island Center, MNR, Shanghai 200135, China)

**Abstract** The Qingcaosha Reservoir in Shanghai is a crucial drinking water source facing the risk of eutrophication. Accurate prediction of chlorophyll-a (Chl-a) concentration is essential for ensuring water safety. This study proposed a graph convolutional temporal pattern attention network (GC-TPA) to predict Chl-a concentration in the reservoir. The model first utilized the temporal pattern attention (TPA) mechanism to capture the temporal dependence of water quality data. It then employed a graph convolutional network (GCN) to learn the relationships between different water quality parameters. In addition, to further improve the prediction accuracy, complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) was introduced to reduce the model's lag, and a multi-layer perceptron (MLP) was used to learn the mutation of Chl-a. The results demonstrated that (1) the GCN module significantly enhanced the Chl-a prediction ability of TPA, and the combination with CEEMDAN and MLP further improved the model performance, with a 56.5% increase in 24-hour prediction accuracy compared to TPA; (2) on a longer prediction period (48 hours),

[收稿日期] 2024-06-04

[基金项目] 上海市级科技重大专项: 人工智能基础理论与关键核心技术 (2021SHZDZX0100)

[作者简介] 彭涛 (1998—), 男, 硕士, 研究方向为水环境模型, E-mail: pts@tongji.edu.cn。

[通信作者] 张海平 (1966—), 男, 教授, 研究方向为水环境模拟与富营养化机制, E-mail: hpzhang@tongji.edu.cn。

GC-TPA still outperformed the TPA and long short-term memory (LSTM) models, with 25.5% and 24.0% lower average absolute and mean square errors, respectively, than LSTM, with 4.92% and 8.40% lower than TPA; and (3) the GCN module improved the feature learning ability of TPA and enhanced the prediction accuracy on the Chl-a trend, while the MLP module was more sensitive to Chl-a mutations. The proposed GC-TPA model performed well when predict of Chl-a concentration in Qingcaosha Reservoir, providing a viable approach for water quality management and ensuring the safety of drinking water.

**Keywords** eutrophication chlorophyll-a (Chl-a) graph convolutional networks (GPN) temporal pattern attention (TPA) multilayer perceptron

饮用水安全关系到广大人民群众的健康,是人类生存的基本需求<sup>[1]</sup>。目前,富营养化仍然是世界上大多数淡水的主要水质问题,也是我国淡水安全面临的重大威胁之一<sup>[2-3]</sup>,湖泊、水库等淡水水体的富营养化正严重威胁我国的饮用水安全。叶绿素 a (chlorophyll-a, Chl-a) 作为评价水体富营养化的重要指标,对其在水体中的浓度变化趋势进行准确预测可为湖库水质管理和应急响应提供有效支撑,对饮用水安全保障具有重要意义。

国内外针对水体 Chl-a 浓度的预测方法已有大量研究。但由于水体富营养化成因复杂,影响因素众多且藻类生长机制尚不十分明确,许多传统的机理模型如 EFDC<sup>[4]</sup>、WASP<sup>[5]</sup> 以及 MIKE<sup>[6]</sup> 等往往难以实现 Chl-a 浓度的准确预测,且这类模型通常存在参数众多、过程复杂、计算量大等问题。而传统的数理统计方法如差分自回归移动平均 (autoregressive integrated moving average, ARIMA) 模型<sup>[7]</sup>、主成分分析 (principal component analysis, PCA)<sup>[8]</sup> 等难以拟合 Chl-a 与其影响因子之间复杂的非线性关系,从而导致预测效果较差。近年来,机器学习、深度学习等技术飞速发展,相关模型已被广泛应用于气象、水文、医学、交通等众多领域<sup>[9-12]</sup>。其中用于时序预测的循环神经网络[如循环神经网络 (RNN)、长短时记忆网络 (long short-term memory, LSTM)、门控循环单元 (gated recurrent unit, GRU) 等] 以及卷积神经网络 (convolutional neural network, CNN) 等在水质预警预报研究中得到大量应用。

水环境系统的复杂性和多样性给使用单一模型预测水质浓度带来了巨大挑战,结合多种模型优势的复合模型逐渐成为研究热点。Liu 等<sup>[13]</sup> 通过结合时空特征对 CNN-LSTM 模型进行改进来预测 Chl-a 浓度,在长期预测中取得了优于其他基线模型的预测效果。丁艺鼎等<sup>[14]</sup> 基于 LSTM 和 GRU 构建了一

个可解释性的框架用于预测湖泊总磷浓度,取得了较好的预测精度。但随着研究<sup>[15]</sup> 的深入,人们发现传统的 RNN 和 CNN 往往难以捕获长距离的时间依赖性。注意力机制被提出以来,已经在计算机视觉、自然语言处理、交通预测以及机器翻译等众多领域取得重大成就<sup>[15-18]</sup>,各种基于注意力机制的时序预测模型相继提出<sup>[19-20]</sup>,为水质预测研究提供了新的方法。Shih 等<sup>[21]</sup> 提出的时间模式注意力 (temporal pattern attention, TPA) 模型在电力消耗、太阳能发电等多元时间序列预测任务中表现出强大的性能。在 TPA 中,使用了一维 CNN 滤波器对循环网络的隐藏状态进行卷积,从而捕获多个时间步长的时间模式。这本质上仍是一种时间卷积,虽然结合后续的打分机制可以让模型学习特征之间的相互依赖性,但该模型的特征学习能力仍有不足。在图卷积网络 (graph convolutional network, GCN) 中,对每个结点特征的学习考虑了其所有邻居以及其自身所包含的特征信息<sup>[22]</sup>,这有助于提升模型的特征学习能力。

本文在 TPA 模型的基础上,引入图学习和图卷积模块替换 TPA 中的一维卷积模块,将时间序列转化为图结构,经过图卷积提取特征后计算注意力分数,最后产生预测结果。此外,在研究中发现,当输入数据中包含 Chl-a 浓度时,模型往往难以学习到其他特征,同时由于时间序列的自相关性导致预测结果存在严重的滞后。另外,水质指标在时间变化上的连续性,Chl-a 浓度的历史观测值在模型输入中不可或缺,因此本文利用 CEEMDAN 算法对 Chl-a 进行分解后作为模型输入,同时增加了一个多层感知机 (multilayer perceptron, MLP) 模块学习 Chl-a 序列的非平稳性,用来弥补数据分解带来的过度平稳。

本研究以上海市青草沙水库为研究对象,该水库具有库容量大、水深浅、水力停留时间较长等特点,同时受上游长江来水水质的影响,面临较大的藻华暴发风险。为了保障饮用水供水安全,对水库未

来一段时间(24~48 h)的水质变化进行准确预测,从而为水库水质管理提供指导显得尤为重要。基于此,本文(1)基于 TPA 和 GCN,引入 CEEMDAN 算法和 MLP 模块,提出了 GC-TPA 模型并验证了所提出改进措施的有效性;(2)通过对比原始 TPA 模型和 LSTM 模型验证了 GC-TPA 模型的预测能力;(3)进一步分析了 MLP 模块所发挥的作用。

## 1 数据与方法

### 1.1 研究区概况与数据采集

青草沙水库位于上海长兴岛西北方冲积沙洲青草沙上[图 1(a)],于 2010 年底建成并开始供水。作为上海市百年战略水源地,维持其优良水质对于保障全市 1 300 万人饮用水安全和维持城市经济健康发展具有重要意义。该水库总面积为 66.26 km<sup>2</sup>,设计有效库容为 4.35 亿 m<sup>3</sup>,设计供水能力为 719 万 m<sup>3</sup>/d,目前平均供水量为 510 万 m<sup>3</sup>/d。青草沙水库上游来水营养盐浓度高,库区实测营养盐总磷平均质量浓度为 0.18 mg/L、总氮平均质量浓度为

1.54 mg/L<sup>[23]</sup>,已经达到水体富营养化的标准(总磷平均质量浓度为 0.03~0.10 mg/L、总氮平均质量浓度为 1.0~2.0 mg/L<sup>[24]</sup>)。另外,该水库地形条件复杂,局部存在缓滞流区,进一步增加了藻华风险。

目前,青草沙水库已建立较为完善的水质监测与数据采集系统,在水库输水口、上游库内、下游库内等几个关键点位每天进行一次人工采样分析[图 1(b)],分析指标包括 Chl-a、pH、溶解氧、高锰酸盐指数、五日生化需氧量、水温、浑浊度、总氮、总磷、氨氮等,除遇特殊情况外,每天的采样时间相对固定,从而保证了数据的一致性。本研究选取了 2017 年 1 月 8 日—2021 年 12 月 31 日共计 1 819 条人工监测数据作为试验样本。此外,藻类生长受到光照等天气因素影响较大,为了更准确地预测水库 Chl-a 变化,采用了美国国家航空航天局的 POWER 数据集(<https://power.larc.nasa.gov/data-access-viewer/>)中的每日气象数据,包括地表净太阳辐射、地表 2 m 大气温度、地表 2 m 风速等。

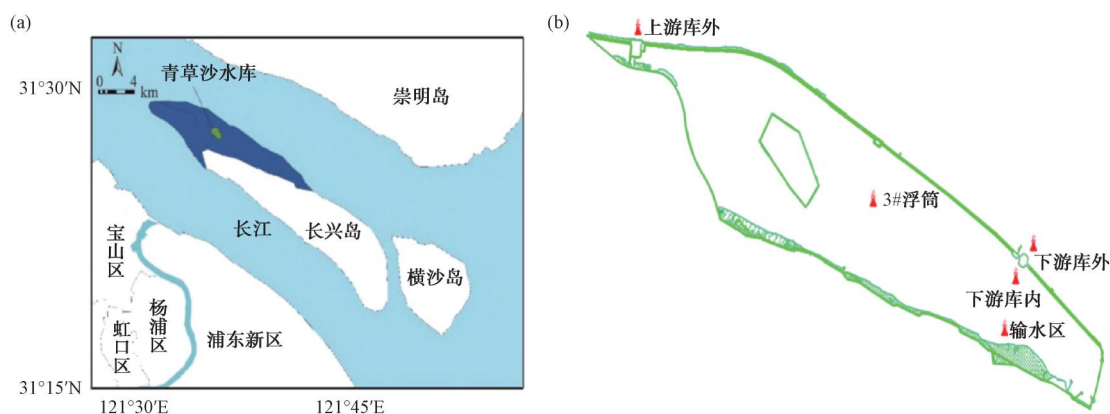


图 1 青草沙水库地理位置和监测点位

Fig. 1 Location and Monitoring Sites of Qingcaosha Reservoir

### 1.2 数据预处理与特征选择

由于获取到的水质和气象数据较多,包含了很多冗余的数据,输入冗余的数据不仅增大了计算量、浪费储存空间,还会导致模型过拟合。因此,有必要进行特征筛选以排除不必要的特征。本文通过皮尔逊相关系数法和单变量选择法分析各参数对叶绿素预测的重要性,皮尔逊相关系数法中采用相关系数的大小作为特征筛选的依据,综合两种方式的结果,最终选取了输水口 Chl-a (Chl-a<sub>ss</sub>)、五日生化需氧量(BOD<sub>5</sub>)、下游库内的 Chl-a、库中央的 Chl-a、10 m 范围内的风速、地表 2

m 处最高温度、地表 10 m 风向、地表净太阳辐射共 8 个指标。此外,氮磷营养盐和水温虽然在相关性分析中未表现出与 Chl-a 的显著相关性,但由于它们对藻类生长的重要影响,经进一步分析后,将氨氮、总磷和水温一并纳入模型输入特征。为减少模型对历史数据的过度依赖造成预测结果的滞后性,Chl-a<sub>ss</sub> 首先经 CEEMDAN 分解为 9 个模态分量信号。最终模型一共有 20 个输入特征。

此外,为了消除各指标之间的量纲影响,提升模型的稳定性和精度,加快模型收敛速度,采用最大最小值归一化法对数据进行归一化,其计算如式(1)。

$$x'_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \quad (1)$$

其中:  $\min(x)$ 、 $\max(x)$ ——样本数据序列的最小值、最大值;

$x_i$ 、 $x'_i$ ——第  $i$  个数据点、其标准化后的值。

### 1.3 模型原理及建模过程

#### 1.3.1 TPA

TPA 的原理如图 2 所示,其中  $h_t$  表示循环神经

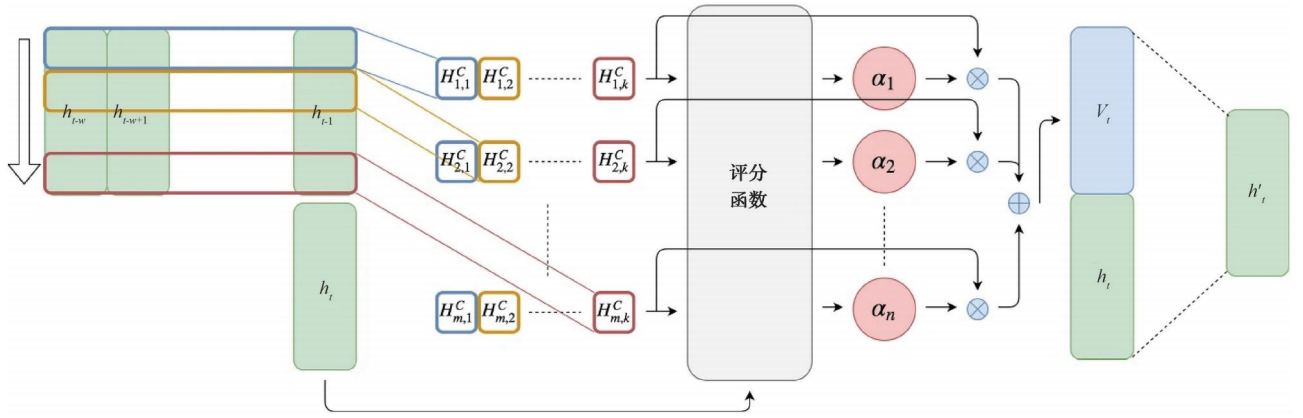


图 2 TPA 原理<sup>[21]</sup>

Fig. 2 Principles of TPA<sup>[21]</sup>

TPA 机制引入了时间模式编码和基于时间模式的注意力计算,使得模型能够更好地理解和捕捉时间序列数据中的重要模式和特征,从而提高预测性能。

#### 1.3.2 GCN

Kipf 等<sup>[22]</sup>提出的 GCN 是传统卷积算法在图结构数据上的一个变体,广泛应用到非欧结构数据的空间特征提取,通过学习图中每个节点与其邻居节点和更远节点之间的相互关系从而得到节点的特征。在本文中,将  $N$  个特征向量在  $t$  时刻的观测值  $x_t \in \mathbb{R}^N$  转化为图的特征矩阵  $X \in \mathbb{R}^{N \times M}$ ,其中  $M$  表示时间序列的长度。不同特征之间的关系可以表示为一张确定的无向图  $G = (V, E, A)$ ,其中  $V$  代表图中的节点向量,  $E$  代表连接节点的边,  $A \in \mathbb{R}^{N \times N}$  代表邻接矩阵,表示节点之间的空间关系。 $L$  层的图卷积过程如式(2)。

$$H^l = \sigma \left( \tilde{D} - \frac{1}{2} \tilde{A} \tilde{D} - \frac{1}{2} H^{l-1} \omega^l \right) \quad (2)$$

其中:  $H^{l-1}$  和  $H^l$ ——输入和输出 ( $H^0 = X \in$

网络在时间步  $t$  的隐藏状态,有  $k$  个长度为  $w$  的一维 CNN 滤波器,用不同颜色的矩形表示。每个滤波器对隐藏状态的  $m$  个特征进行卷积,并生成具有  $m$  行和  $k$  列的矩阵  $H^C$ 。评分函数通过与当前隐藏状态  $h_t$  进行比较来计算  $H^C$  每行的权重。对权重进行归一化,并将  $H^C$  的行与其相应的权重进行加权求和以生成  $V_t$ 。最后,连接  $V_t$ 、 $h_t$  并执行矩阵乘法以生成  $h'_t$ ,用于生成最终的预测结果。

$\mathbb{R}^{N \times M}$ );

$\tilde{A}$  和  $\tilde{D}$ ——修正后的邻接矩阵及其度矩阵;

$\sigma$ ——非线性激活函数;

$\omega^l$ —— $l$  层的可学习参数。

由于水质指标之间的依赖关系并不明确,使得邻接矩阵  $A$  的构建成为一个难点。为此, Wu 等<sup>[25]</sup>提出了图学习模块自动生成图结构的邻接矩阵,主要计算过程如式(3)~式(7)。

$$M_1 = \tanh(\alpha E_1 \Theta_1) \quad (3)$$

$$M_2 = \tanh(\alpha E_2 \Theta_2) \quad (4)$$

$$A = \text{ReLU} \{ \tanh[ \alpha (M_1 M_2^T - M_2 M_1^T) ] \} \quad (5)$$

for  $i = 1, 2, \dots, N$ ;

$$\text{index} = \text{argtopk}(A[i, :]) \quad (6)$$

$$A[i, -\text{index}] = 0 \quad (7)$$

其中:  $E_1$ 、 $E_2$ ——随机初始化的节点嵌入,在模型训练过程中是可学习的;

$\Theta_1$ 、 $\Theta_2$ ——模型参数;

$\alpha$ ——用于控制激活函数饱和率的超



参数;

$A[i, :]$ ——矩阵  $A$  的第  $i$  行。

argtopk 函数返回向量最大的  $k$  个值。通过式 (6) 和式 (7) 使邻接矩阵变得稀疏,从而减少后续图卷积的计算成本。

### 1.3.3 GC-TPA 模型

文章提出的 GC-TPA 模型结构如图 3 所示。Chl-a<sub>ss</sub> 数据经 CEEMDAN 分解后,与地表净太阳辐射、气温等其他输入特征一起作为 TPA 的输入。数据分解减小了模型预测结果的滞后性,提高了预测精度,但数据分解使模型趋于平稳,因此在 Chl-a 变化较大时精度会降低。为了解决这个问题,在 TPA 之外增加了一个包含两个线形层的 MLP 模块,该模块以  $t$  时间步的其他特征作为输入,用以学习 Chl-a 的突变。经过反复试验,发现将两部分模型的预测结果直接相加可以使得模型在训练集和测试集上的表现最为均衡,因此在 GC-TPA 中采用直接相加的方式得到  $t+1$  时间步的预测结果  $y_{t+1}$ 。

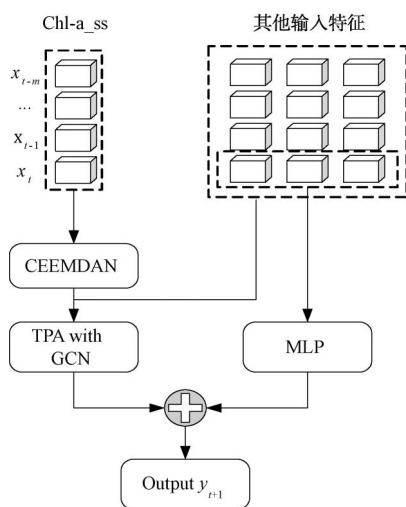


图 3 GC-TPA 模型整体架构

Fig. 3 Architecture of GC-TPA Model

### 1.4 模型性能评价

为了验证提出的复合模型的效果,文章选择了 LSTM 模型与原始 TPA 模型作为基线模型。选择均方根误差 (mean square error, RMSE)、平均绝对误差 (mean absolute error, MAE) 和纳什效率系数 (nash-sutcliffe efficiency coefficient, NSE) 作为模型性能的评价标准。MAE 表示预测值与真实值之间的绝对误差的均值,可以有效地反映预测误差;RMSE 表示预测值与真实值之间的离差,可以衡量

预测值中是否存在异常值;NSE 表示预测值与真实值之间的一致性。其中 MAE 和 RMSE 越小越好, NSE 越接近 1 表示模型预测值与真实值越接近。其计算方法如式 (8) ~ 式 (10)。

$$M_{(y, \hat{y})} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (8)$$

$$R_{(y, \hat{y})} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (9)$$

$$N_{(y, \hat{y})} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (10)$$

其中:  $M_{(y, \hat{y})}$ ——MAE;

$R_{(y, \hat{y})}$ ——RMSE;

$N_{(y, \hat{y})}$ ——NSE;

$y_i$ ——第  $i$  个真实值;

$\hat{y}_i$ ——第  $i$  个预测值;

$\bar{y}$ ——序列  $y$  的平均值;

$n$ ——序列长度。

## 2 结果与分析

### 2.1 CEEMDAN 分解

如前文所述,为了降低模型对 Chl-a 数据的依赖性,对 Chl-a<sub>ss</sub> 数据进行了 CEEMDAN 分解,图 4 为不包括残差项的分解结果,分别为 IMF1 ~ IMF9 共 9 个模态分量。可以看出,IMF1 ~ IMF7 为 Chl-a 数据的高频分量,反映的是 Chl-a 的短期变化;IMF8 和 IMF9 为 Chl-a 数据的趋势分量,反映的是 Chl-a 的长期变化趋势。从长期来看,Chl-a 浓度呈现出先上升后下降的趋势,而它的短期变化则相对复杂,难以发现规律,这对模型的预测而言是一个难点。因此才需要增加其他输入特征,以增强模型的预测能力。

### 2.2 消融试验

为了验证 GCN 模块和 MLP 模块的有效性,文章设计了一个单步预测的消融试验。通过从模型中移除掉不同的模块,然后分别采用控制变量的方法研究这些模块对于模型性能的影响<sup>[26]</sup>。试验结果如表 1 所示。

消融试验的结果表明 MLP 模块和 GCN 模块的加入都发挥了积极作用。使用 GCN 替换 TPA 的一维卷积后模型性能得到了显著提升,各项评价指标相较于 TPA 都得到了提高,尤其是纳什系数从 0.249 0 提升到了 0.308 3,预测值与真实值之间的

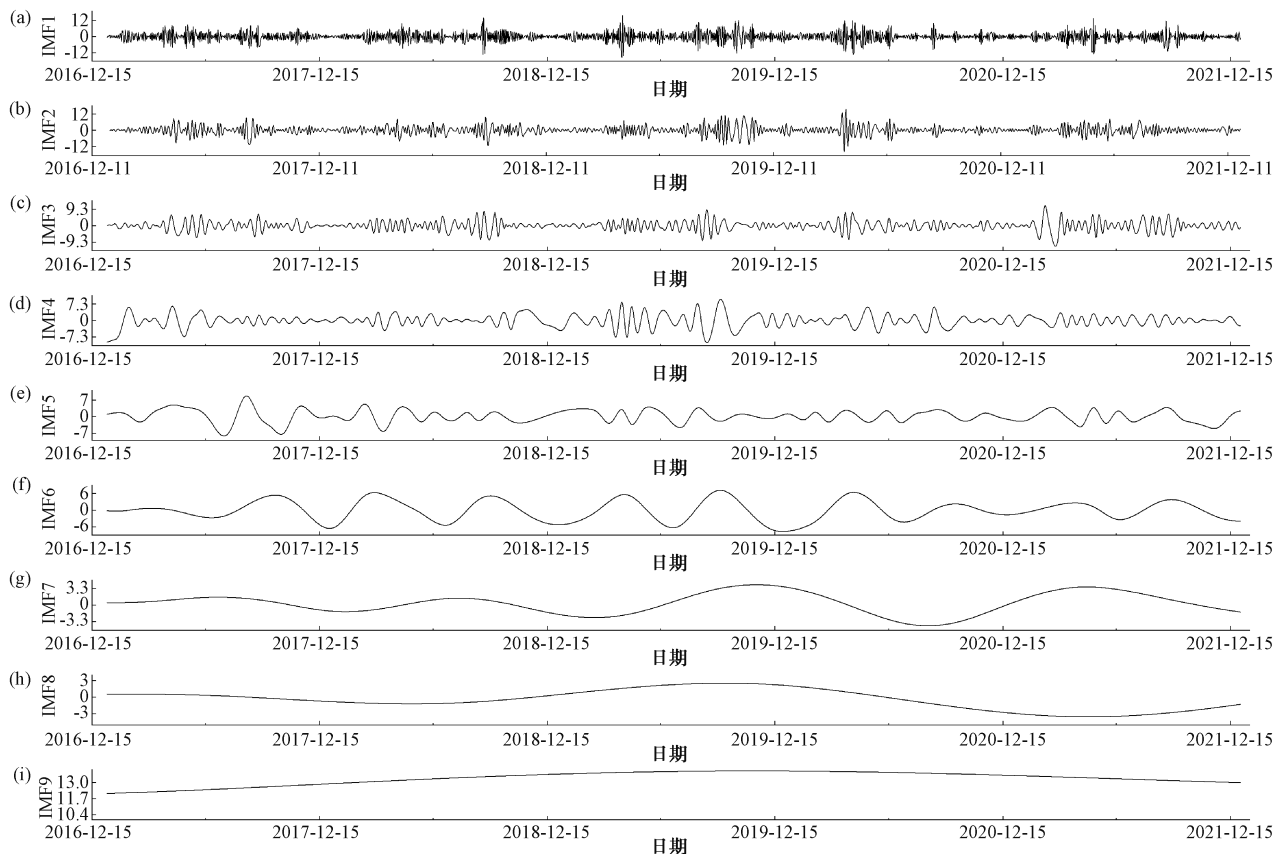


图 4 Chl-a 数据分解结果

Fig. 4 Chl-a Data Decomposition Results

表 1 消融试验结果

Tab. 1 Results of Ablation Experiments

模型	MAE	RMSE	NSE
TPA	3.552	4.746	0.249 0
TPA+GCN	3.395	4.570	0.308 3
TPA+MLP	3.634	4.919	0.190 8
GC-TPA	3.328	4.212	0.389 6

接近程度有了很大的提升。TPA+MLP 的组合模型性能相比 TPA 有所降低,是由于 TPA 的特征学习能力较弱,MLP 的加入将导致模型输出偏向于前一天中其他站点的 Chl-a 数据,导致了性能下降。但 MLP 和 GCN 的共同作用却使模型的性能进一步得到提升,在各个评价指标上都优于原始的 TPA 模型,也优于 TPA+GCN 的组合,可见 GCN 的加入显著提高了模型的学习能力。综合来看,结合了 TPA、GCN 和 MLP 的 GC-TPA 模型在各项评价指标上都取得了最优的性能表现,消融试验结果表明混合网络可以充分发挥各个模块的优势,提高模型的

预测精度,从而证明了这些模块的有效性。

另外,为了验证数据分解是否有效减轻了模型对 Chl-a 历史数据的依赖性并提升模型性能,使用未分解的 Chl-a 数据重新对 GC-TPA 模型进行训练和测试,模型在测试集上的表现如图 5 所示。由图 5 可知,在不进行数据分解的情况下,模型的预测值与真实值相比存在大约 2 d 的滞后,并不能很好的实现预测任务。且未进行数据分解时,模型在测试集上的各项评价指标如下。MAE 为 4.347, RMSE 为 5.531, NSE 为 -0.145 6, NSE<0 表明模型已经失去了可信度。

### 2.3 模型性能评估与讨论

通过多次试验和尝试,最终确定取得最优预测精度的 GC-TPA 模型的主要结构参数和超参数如表 2 所示。

为了评估所提出的 GC-TPA 模型在 Chl-a 预测上的实际表现,使用前文中筛选出来的特征作为输入,Chl-a<sub>ss</sub> 作为预测目标。采用输入长度为 7 d,输出长度为 1 d 的模式,建立 GC-TPA、TPA 和 LSTM

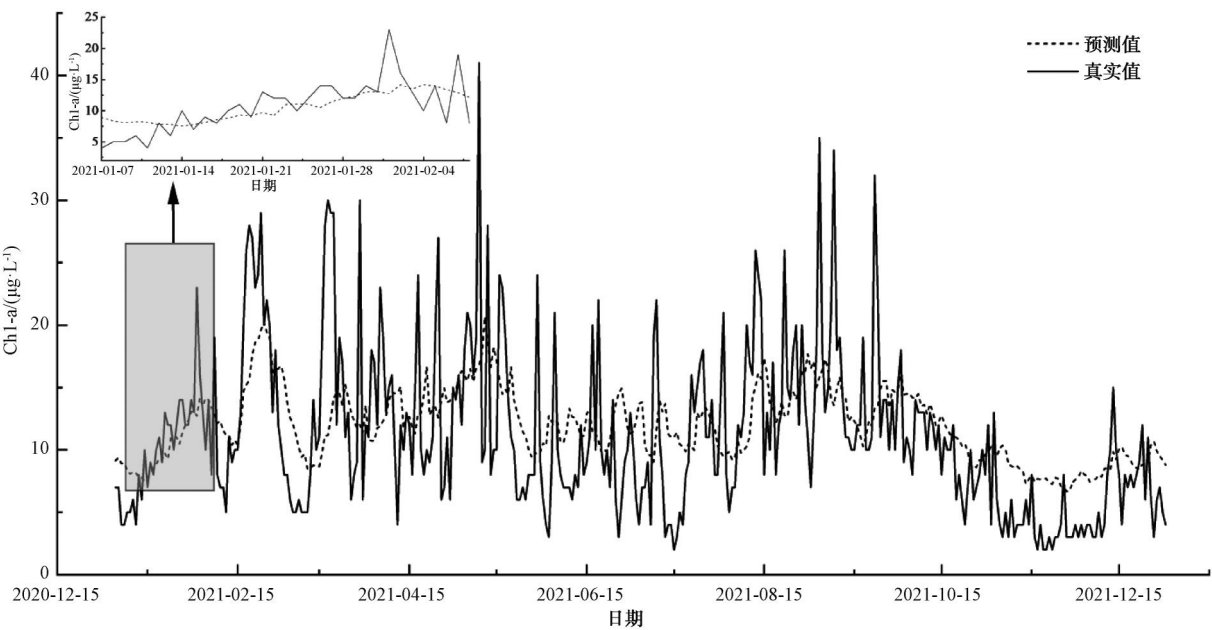


图 5 未进行数据分解时 GC-TPA 模型的测试集表现  
Fig. 5 Test Set Performance of GC-TPA Model without Data Decomposition

表 2 GC-TPA 模型结构

Tab. 2 Structure of GC-TPA Model

参数类型	层级	项目	值
模型结构	输入层	—	20×7
	TPA 模块	—	—
	LSTM 层	单元数	32
	GCN 层	输出通道数	3
	Attention	输出通道数	32
	输出	—	1×1
	MLP 模块	—	—
	输入	—	11×1
	单元数	—	32
	输出	—	1×1
超参数	初始学习率	—	0.001
	批处理大小	—	32
	最大批次	—	30
	Dropout	—	0.2

模型并分别以 24 h 和 48 h 后的 Chl-a 浓度作为预测目标进行训练,并在测试集上评估模型性能,试验结果如表 3 所示。

由表 3 可知,所提出的 GC-TPA 模型在 Chl-a 预测方面的性能优于 TPA 和 LSTM 模型。其中 LSTM 模型的 MAE 和 RMSE 虽然与其他两个模型相差不大,但其 NSE 始终为负数,表明该模型完全没有学

表 3 GC-TPA 与基线模型的性能比较

Tab. 3 Performance of GC-TPA Compared with Baseline Models

模型	24 h			48 h		
	MAE	RMSE	NSE	MAE	RMSE	NSE
TPA	3.552	4.746	0.249 0	3.780	5.152	0.126 1
LSTM	4.994	6.246	-0.710 9	4.822	6.213	-0.569 7
GC-TPA	3.328	4.212	0.389 6	3.594	4.719	0.265 0

习到数据的趋势和特征。在未来 24 h 和 48 h 的 Chl-a 预测任务中,各项评价指标均取得了最好成绩,尤其是在表示模型预测值与真实值一致性的 NSE 上,新模型相比其原型提升了 56.5%。相较于预测 24 h 后的 Chl-a 浓度,在预测 48 h 后的 Chl-a 浓度时模型的性能出现下降,但 GC-TPA 仍保持了不错的性能,在部分评价指标上的表现甚至优于 TPA 的 24 h 预测性能。

3 个模型在测试集上的 24 h 预测结果如图 6 所示。可以看出,TPA 和 GC-TPA 对于预测 Chl-a 的整体变化趋势都有较好的表现,但是当 Chl-a 浓度变化较大,尤其是在峰谷处,GC-TPA 具有更快的响应速度和更高的预测精度,也因此 GC-TPA 的预测结果与实测值的一致性更好,NSE 值更高。当输入数据的时间步长比较长时,LSTM 的预测结果会趋于平稳,再加上数据分解增加了输入数据的平稳性,导致 LSTM 的预测结果则保持在 11 µg/L 波动,未

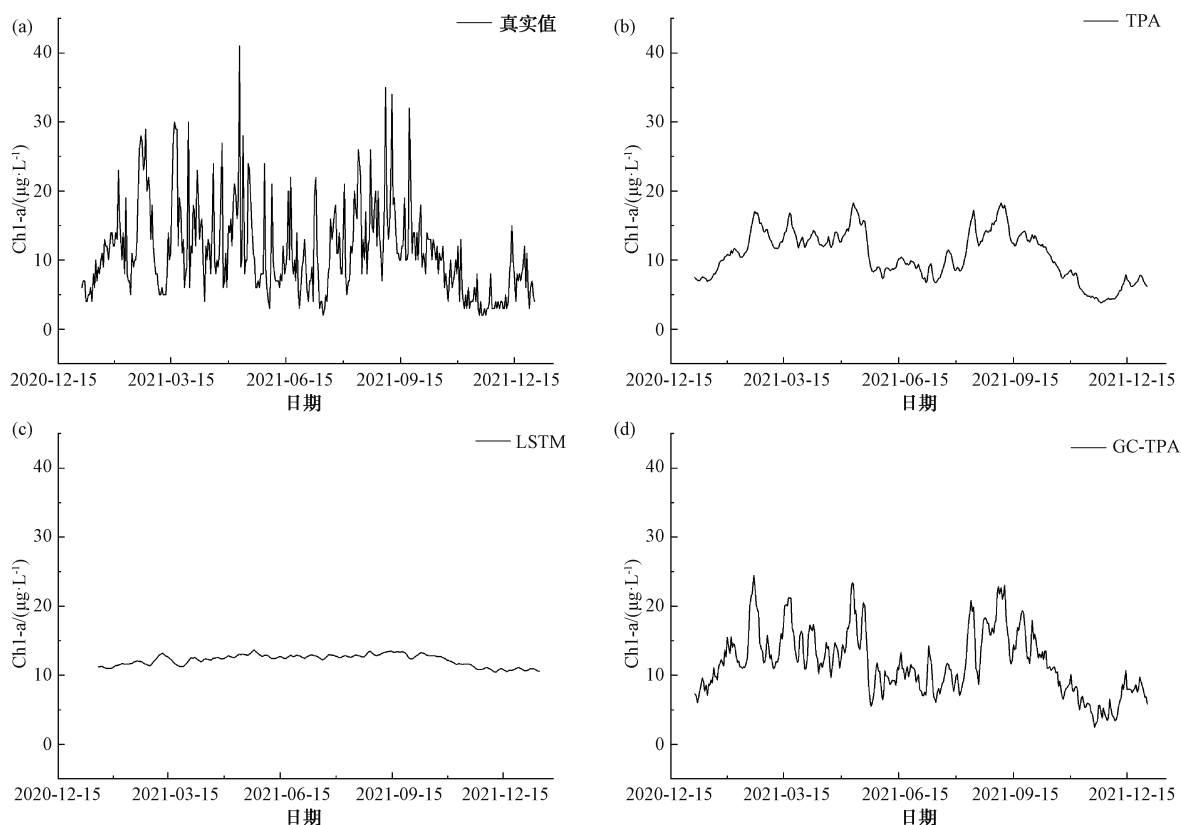


图6 不同模型对未来 24 h Chl-a 浓度的预测值与实测值对比

Fig. 6 Comparison of Predicted and Measured Chl-a Concentrations after 24 h by Different Models

能准确预测 Chl-a 的变化趋势。

图 7(a) 为 GC-TPA 模型中 TPA 模块和 MLP 模

块的输出数据。与设计的目的一样, MLP 模块在 Chl-a 变化较小时比较平稳, 在  $5 \mu\text{g/L}$  波动, 但是当

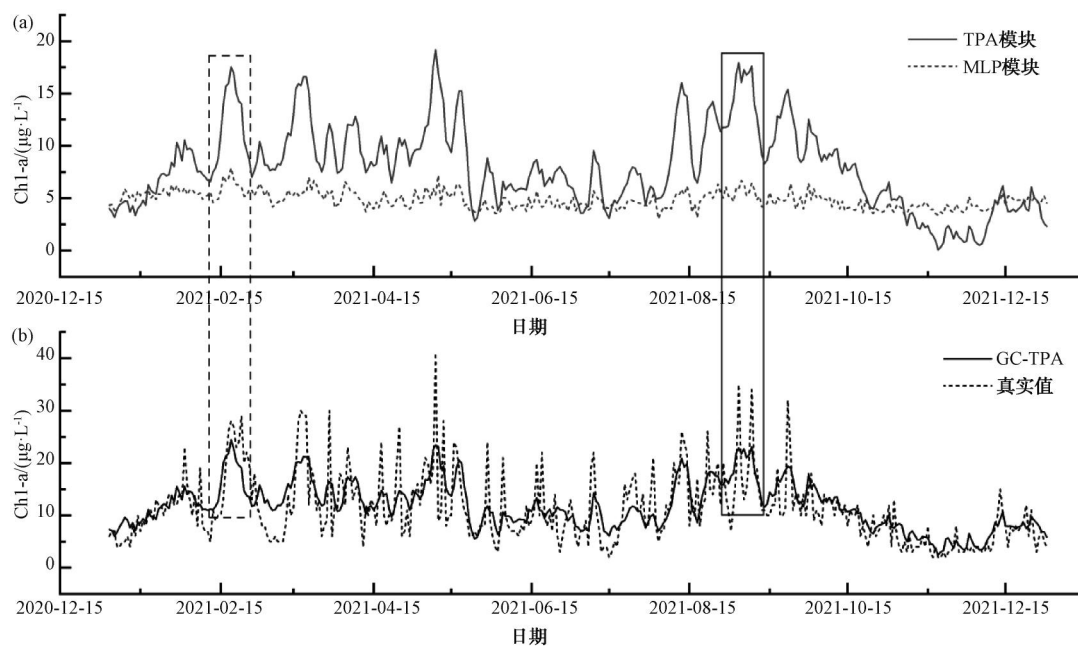


图7 GC-TPA 中两个模块的 24 h 预测输出

Fig. 7 24 h Predicted Output of Two Modules in GC-TPA



Chl-a 浓度出现较大变化时,例如图中虚线框中的部分,MLP 会做出快速响应并弥补 GC-TPA 的不足。如图 3 所示,MLP 模块使用前 1 d 的其他特征作为输入,而水体中 Chl-a 浓度变化是一个连续过程,前 1 d 的特征往往包含了其变化趋势的信息,但由于数据分解使 GC-TPA 模型学习到的趋势更加趋于平稳。因此难以对 Chl-a 浓度的快速变化做出响应,而不使用分解后的 Chl-a 数据作为输入的 MLP 则可以在不损失模型精度的前提下以较小的计算量弥补这一不足。

两个模块的协同使模型能够较好的预测 Chl-a 浓度的变化[图 7(b)],但是模型仍存在一些不足之处,如图中实线框中的部分,当 Chl-a 浓度在短期内出现较大波动时,模型难以快速作出反应,导致对 Chl-a 浓度的预测精度较低。富营养化成因复杂,水中 Chl-a 浓度变化受到多种因素共同影响,本文提出的模型仅考虑了水质和气象因素,因此仍有进一步提升的空间。

### 3 结论

本文基于青草沙水库水质监测数据和气象数据,提出了青草沙水库 Chl-a 预测的 GC-TPA 混合模型。其中,TPA 用于学习水质数据的时间依赖性;GCN 学习不同水质参数之间的关系,提高模型的特征学习能力;MLP 则学习 Chl-a 的突变。混合模型充分发挥了不同模块的优势,提高了 Chl-a 浓度的预测精度。消融试验结果表明,GCN 模块和 MLP 模块的加入显著提升了模型性能,CEEMDAN 数据分解也发挥了积极作用,降低了模型输出的滞后性,提升了预测精度。在与 TPA 和 LSTM 两个基线模型的对比试验中,本研究提出的 GC-TPA 模型在 24 h 和 48 h 的 Chl-a 浓度预测上均表现出更好的性能,其中 24 h 预测的 MAE、RMSE 和 NSE 分别为 3.328、4.212 和 0.389 6;48 h 预测任务中模型精度虽有所下降,但仍表现出比 TPA 模型 24 h 预测更好的性能。通过分析混合模型中 TPA 和 MLP 两个模块的输出,发现 MLP 模块在 Chl-a 浓度快速变化时做出响应,以修正 TPA 模块输出过于平稳的缺陷。本文提出的 GC-TPA 混合模型在青草沙水库 Chl-a 预测中表现良好,可为水库水质管理提供支撑。

### 参考文献

[1] 曹晓峰,冀泽华,兰华春,等. 气候变化背景下我国湖库型

水源富营养化控制与饮用水安全保障策略[J]. 中国工程科学, 2022, 24(5): 34-40.

CAO X F, JI Z H, LAN H C, et al. Eutrophication control of lake/reservoir water sources and assurance of drinking water safety in china considering global climate change[J]. Strategic Study of CAE, 2022, 24(5): 34-40.

[2] HO J C, MICHALAK A M, PAHLEVAN N. Widespread global increase in intense lake phytoplankton blooms since the 1980s[J]. Nature, 2019, 574(7780): 667-670.

[3] WU B, DAI S, WEN X, et al. Chlorophyll-nutrient relationship changes with lake type, season and small-bodied zooplankton in a set of subtropical shallow lakes[J]. Ecological Indicators, 2022, 135: 108571. DOI: 10.1016/j.ecolind.2022.108571.

[4] KIM J, LEE T, SEO D. Algal bloom prediction of the lower Han River, Korea using the EFDC hydrodynamic and water quality model[J]. Ecological Modelling, 2017, 366: 27-36. DOI: 10.1016/j.ecolmodel.2017.10.015.

[5] SEO D, KIM M, AHN J H. Prediction of chlorophyll-a changes due to weir constructions in the nakdong river using EFDC-WASP modelling[J]. Environmental Engineering Research, 2012, 17(2): 95-102.

[6] 张月婷,徐明德,原文超. 基于 MIKE 21 的汾河水库挟沙水流水质数值模拟研究[J]. 四川环境, 2022, 41(1): 147-153.

ZHANG Y T, XU M D, YUAN W C. A numerical simulation study of sand-carrying water quality of fenhe reservoir based on MIKE21 model [J]. Sichuan Environment, 2022, 41(1): 147-153.

[7] 刘庄,汪永国,丁程成,等. 基于 ARIMA 模型的长潭水库叶绿素 a 浓度预测[J]. 环境污染与防治, 2023, 45(7): 895-902.

LIU Z, WANG Y G, DING C C, et al. Prediction of chlorophyll a concentration in Changtan Reservoir based on ARIMA model [J]. Environmental Pollution & Control, 2023, 45(7): 895-902.

[8] 陈昭明,王伟,赵迎,等. 改进主成分分析与多元回归融合的汉丰湖水水质评估及预测[J]. 环境监测管理与技术, 2020, 32(4): 15-19.

CHEN Z M, WANG W, ZHAO Y, et al. Evaluation and prediction of water quality in hanfeng lake based on improved principal component analysis and multivariate regression model [J]. The Administration and Technique of Environmental Monitoring, 2020, 32(4): 15-19.

[9] BI K, XIE L, ZHANG H, et al. Accurate medium-range global weather forecasting with 3D neural networks[J]. Nature, 2023, 619(7970): 533-538.

[10] YAO Z, WANG Z, WANG D, et al. An ensemble CNN-LSTM and GRU adaptive weighting model based improved sparrow search algorithm for predicting runoff using historical

- meteorological and runoff data as input [J]. *Journal of Hydrology*, 2023, 625: 129977. DOI: 10.1016/j.hydrol.2023.129977.
- [11] HAUG C J J, DRAZEN J M M. Artificial intelligence and machine learning in clinical medicine[J]. *New England Journal of Medicine*, 2023, 388(13): 1201–1208.
- [12] ZHANG X, WEN S, YAN L, et al. A hybrid-convolution spatial-temporal recurrent network for traffic flow prediction[J]. *Computer Journal*, 2024, 67(1): 236–252.
- [13] LIU N, CHEN S Y, CHENG Z Y, et al. Long-term prediction of sea surface chlorophyll-a concentration based on the combination of spatio-temporal features [J]. *Water Research*, 2022, 211: 118040. DOI: 10.1016/j.watres.2022.118040.
- [14] 丁艺鼎, 范宏翔, 徐力刚, 等. 可解释性长短期记忆模型用于预测湖泊总磷浓度变化[J]. *湖泊科学*, 2024, 36(4): 1046–1059.
- DING Y D, FAN H X, XU L G, et al. The interpretable long-term and short-term memory model was used to predict the change of total phosphorus concentration in lakes[J]. *Journal of Lake Sciences*, 2024, 36(4): 1046–1059.
- [15] SUBAKAN C, RAVANELLI M, CORNELL S, et al. Attention is all you need in speech separation [C]. Toronto: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.
- [16] ITTI L, KOCH C, NIEBUR E. A model of saliency-based visual attention for rapid scene analysis [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, 20(11): 1254–1259.
- [17] 张文娟, 杨皓哲, 张彬, 等. 考虑多时间尺度特征的城市轨道交通短时客流量预测模型[J]. *交通运输系统工程与信息*, 2022, 22(6): 212–223.
- ZHANG W J, YANG H Z, ZHANG B, et al. Short-time passenger flow prediction model of urban rail transit considering multi-timescale features [J]. *Journal of Transportation Systems Engineering & Information Technology*, 2022, 22(6): 212–223.
- [18] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [C]. San Diego: 3rd International Conference on Learning Representations, 2015.
- [19] ZHOU H, ZHANG S, PENG J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35: 11106–11115. DOI: 10.1609/aaai.v35i12.17325.
- [20] WANG W, CHEN W, QIU Q, et al. CrossFormer ++: A versatile vision transformer hinging on cross-scale attention[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024, 46(5): 3123–3136.
- [21] SHIH S Y, SUN F K, LEE H Y. Temporal pattern attention for multivariate time series forecasting [J]. *Machine Learning*, 2019, 108(8/9): 1421–1441. DOI: 10.1007/s10994-019-05815-0.
- [22] KIPF T, WELLMING M. Semi-supervised classification with graph convolutional networks [C]. Toulon: 5th International Conference on Learning Representations, 2017.
- [23] 刘书敏, 赵风斌. 长江口水源地青草沙水库水质与浮游植物群落特征 [J]. *环境污染与防治*, 2022, 44(10): 1330–1335.
- LIU S M, ZHAO F B. Water quality and phytoplankton community characteristics of a water source in Yangtze Estuary, Qingcaosha Reservoir [J]. *Environmental Pollution & Control*, 2022, 44(10): 1330–1335.
- [24] BHAGOWATI B, AHAMAD K U. A review on lake eutrophication dynamics and recent developments in lake modeling [J]. *Ecohydrology & Hydrobiology*, 2019, 19(1): 155–166.
- [25] WU Z, PAN S, LONG G, et al. Connecting the dots: Multivariate time series forecasting with graph neural networks [C]. New York: 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2020.
- [26] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]. Columbus: 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.