J. Lake Sci.(湖泊科学),2023,**35**(2):449-459
DOI 10.18307/2023.0206

基于机器学习的水—气界面 CO_2、CH_4扩散通量预测及影响因素分析——以三峡水库为例*

欧阳常悦,秦　宇**,刘　臻,梁　越
(重庆交通大学河海学院,环境水利工程重庆市工程实验室,重庆 400074)

摘　要:传统的水—气界面温室气体通量的监测方法具有诸多局限,对其影响因素的分析也大多基于数学统计层面。对此,本研究提供了一种较为新颖的研究和分析方法——基于机器学习的数据预测和分析。本研究采用 2 种经典机器学习算法——随机森林(RF)和支持向量机(SVM)和 2 种深度学习算法——卷积神经网络(CNN)和长短时记忆神经网络(LSTM),通过环境因素预测水库水—气界面 CO_2 和 CH_4 扩散通量。此外,采用 RF 中的特征重要性评估和经典算法决策树(DT),对环境因素和水库温室气体扩散通量的关系进行了全新角度的数据挖掘和分析。结果表明:深度学习算法的预测效果均较好,经典机器学习算法中 RF 预测效果显著优于 SVM。LSTM 和 RF 分别产生了最优的 CO_2 扩散通量和 CH_4 扩散通量的预测精度,均方根误差(*RMSE*)分别为 0.424 mmol/(m^2·h)和 0.140 μmol/(m^2·h),预测值与实测值的 R^2 分别为 0.960 和 0.758。RF 的特征重要性评估表明沉积物因子和营养因子均为影响 CO_2 和 CH_4 扩散通量的关键因子,气候因子和水环境因子相较次之。采用决策树描绘决定 CO_2 扩散通量源和汇的环境因子的极限阈值,决策树对所有样本的分类准确性高达 100%,且其结果还表明低浓度的溶解无机碳和碱性条件有利于水体成为 CO_2 汇。因此,使用机器学习算法预测和分析水库水—气界面温室气体通量的潜力巨大。
关键词:机器学习;深度学习;温室气体通量;预测;三峡水库

Prediction of CO_2, CH_4 diffusion fluxes at the water-air interface and analysis on its influencing factors using machine learning algorithms in the Three Gorges Reservoir*

Ouyang Changyue, Qin Yu**, Liu Zhen & Liang Yue
(*Key Laboratory of Hydraulic and Waterway Engineering of the Ministry of Education, Chongqing Jiaotong University, Chongqing* 400074, *P.R.China*)

Abstract: Traditional methods for monitoring greenhouse gas fluxes at the water-air interface in reservoirs have many limitations. The analysis on its influencing factors is also mainly based on mathematical statistics. This study provides an innovative approach by using machine learning algorithms. In this study, two traditional machine learning algorithms (Random forests (RF) and Support vector machine (SVM)) and two deep learning algorithms (Convolutional neural network (CNN) and long and short term memory neural network (LSTM)) were applied to predict CO_2 and CH_4 diffusion fluxes. In addition, the feature importance assessment in RF and the decision tree (DT) are used to analyze the relationship between environmental factors and GHG diffusion fluxes in reservoirs from a new perspective. The results showed that deep learning produced excellent prediction accuracy, whereas prediction accuracy of RF was significantly better than SVM in traditional machine learning. LSTM and RF yielded optimal accuracy in predicting CO_2 flux and CH_4 flux, respectively. The root mean square error (*RMSE*) was 0.424 mmol/(m^2·h) and 0.140 μmol/(m^2·h) and R^2 of the predicted and measured values were 0.960 and 0.758, respectively. RF identified sediment and nutrient as critical environmental factors to GHG fluxes, followed by climate factors and water environment factors. Lastly, a decision tree was used innovatively to depict the limiting threshold of environmental factors that determines the source or sink of CO_2. The classification accuracy of this decision tree is as high as 100% in this study. The results of decision tree also showed that low dissolved inorganic

* 2022-10-20 收稿;2022-12-24 收修改稿。
国家自然科学基金项目(51609026)和重庆市研究生科研创新项目(CYS22402)联合资助。
** 通信作者; E-mail:qinyu@cqjtu.edu.cn。

carbon concentration and alkaline conditions are favorable for water to absorb atmospheric CO_2. These results demonstrate the great potential of using machine learning algorithms to predict and analyze GHG fluxes at the water-air interface in reservoirs.

Keywords: Machine learning; deep learning; greenhouse gas flux; prediction; Three Gorges Reservoir

全球变暖是世界各国关注的环境问题,各国政府都对此采取积极行动。中国政府提出在 2030 年前实现碳达峰,在 2060 年前实现碳中和。水库是内陆水生态系统的重要组成部分,也是温室气体的重要来源[1-3]。现有研究表明全球水库 CO_2和 CH_4排放量分别为 48.0 Gg/a 和 13.3 Gg/a[4]。然而,水库水—气界面温室气体通量的强烈时空异质性导致通量估算的不确定性[5-6]。此外,监测水库水—气界面温室气体通量的常用方法(例如薄边界模型、静态箱法和倒置漏斗法等)虽各有其适用范围和优点,但也都存在采样周期长、成本高、受瞬时流速和风速影响大、需要设置大量采样点等共同的局限。准确估算内陆水体温室气体排放是预测气候变化的重要前提,也是实现碳达峰和碳中和的重要基础[7]。以往的研究常采用数学统计模型来预测水库温室气体排放。例如一些研究基于热力学平衡理论,采用 pH 值、溶解无机碳(DIC)、碱度等与 CO_2通量的线性回归关系来估算湖泊和水库中的 CO_2通量,但预测通量比实测值几乎低一半[8-10]。或者建立温室气体排放量与生产力(叶绿素 a、总磷)的统计模型,并基于遥感测量的叶绿素 a 来预测全球水库的温室气体排放量[11]。

随着计算机算法的大力发展,机器学习作为人工智能的一个分支,和一种全新的数据挖掘和模型开发的方法出现[12]。它可以洞察大量输入特征与输出之间潜在的相互作用从而实现预测输出。这种方法不同于数学模型的复杂的相互作用,且一般具有更加优良的预测性能[13]。在经典机器学习的分类和回归中,支持向量机(SVM)和随机森林(RF)通常被用于实现这种预测。此外,深度学习是机器学习的一个更加前沿而复杂的新分支,它由深度神经网络驱动,可以模仿人脑并分析文本、图像和音频等数据。深度神经网络能够从原始数据中发现并学习规律,并高效地执行特征预测和分类[14]。机器学习技术在预测水生系统中温室气体通量方面已显示出其一定的优势。Chen 等[15]建立了 8 个神经网络,并研究了其在美国 quoit 湾和 massachusetts 沿海盐沼中预测温室气体通量的适用性,最终表明径向基函数神经网络(RBNN)具有最高的预测精度。Mosher 等[16]研究表明 RF 分别解释了 22.7% 和 20.9% 的 CO_2和 CH_4扩散变化,并将 pH、温度和溶解氧确定为 Douglas 水库最有价值的预测因子。Hyungseok 等[10]研究表明,使用电导率、溶解氧和总溶解性碳的 RF 预测模型具有最优的温室气体通量预测性能,其均方根误差(*RMSE*)为 749.4 mg CO_2/(m^2·d),R^2 = 0.844。目前,水库水—气界面的温室气体排放已受到我国学者的广泛关注和研究[17-18],然而国内鲜有利用机器学习(特别是深度学习)预测水生态系统水—气界面温室气体排放的研究。此外,现有的大量研究主要采用数学统计类方法分析温室气体通量的影响因素,例如应用最广泛的 Spearman 相关性分析[19],且据我们所知,目前还没有研究提供可视化的温室气体源汇分类模型。

为了证实机器学习算法预测水库水—气界面温室气体通量的可行性和准确性,本研究以三峡水库为例,使用经典机器学习算法(RF 和 SVM)和深度学习算法(CNN 和 LSTM),挖掘环境因素与气体通量之间的线性或非线性关系,并预测扩散通量。此外,本研究采用 RF 的特征重要性评估功能来量化环境因子的重要性。最后,采用决策树来可视化地区分 CO_2扩散通量的源汇。本研究提供了一种利用机器学习算法预测和分析水库水—气界面温室气体通量的方法,旨在将先进的人工智能算法引入到该领域未来的研究和实践中。

1 材料与方法

1.1 原始数据

选择三峡水库中段为研究区,于 2019 年 4—9 月逐月监测水体 CO_2和 CH_4的溶存浓度及 12 个环境因素,并采用薄边界层模型(TBL)估算 CO_2和 CH_4的水—气界面扩散通量。监测地点为三峡水库中段的万州、高阳和黄石,每个地点采集 3 个相关样本。共计获得 54 组有效样本,具体的采样点分布、监测方法、计算方法和所有数据详见文献[20]。

前期研究结果表明:CO_2扩散通量变化范围为 -0.499 ~ 9.185 mmol/(m^2·h),其平均值为(2.487 ± 2.010) mmol/(m^2·h)。7 月的 6 个样本表现为 CO_2的汇(通量<0),而其余 48 个样本表示 CO_2的源(图 1a)。

研究期内 CH_4扩散通量为(1.258±1.301) μmol/(m^2·h),在0.129~6.850 μmol/(m^2·h)范围内波动,且所有样本均为 CH_4的源(图1b)。此外,监测的12个环境因子包括水温(WT)、溶解氧(DO)、pH、溶解有机碳(DOC)、溶解无机碳(DIC)、氨氮(NH_3-N)、硝态氮(NO_3^--N)、溶解总氮(DTN)、溶解总磷(DTP)、沉积物总有机碳(TOCs)、沉积物总氮(TNs)、沉积物总磷(TPs)。Pearson 相关性分析结果表明,CO_2扩散通量与 WT、DO 和 pH 呈显著负相关,与 DOC、DIC、NH_3-N、NO_3^--N 和 DTN 呈显著正相关(P<0.05)。而 CH_4扩散通量与环境因子均无显著相关性(图2)。

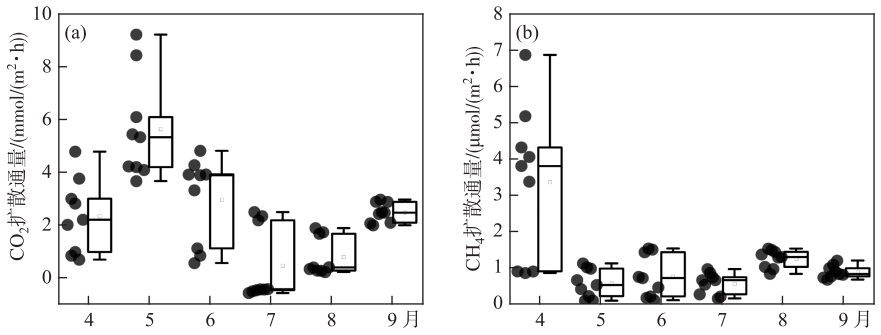

图1 研究期内实测 CO_2和 CH_4扩散通量

Fig.1 Actual CO_2 and CH_4 diffusion fluxes in study periods

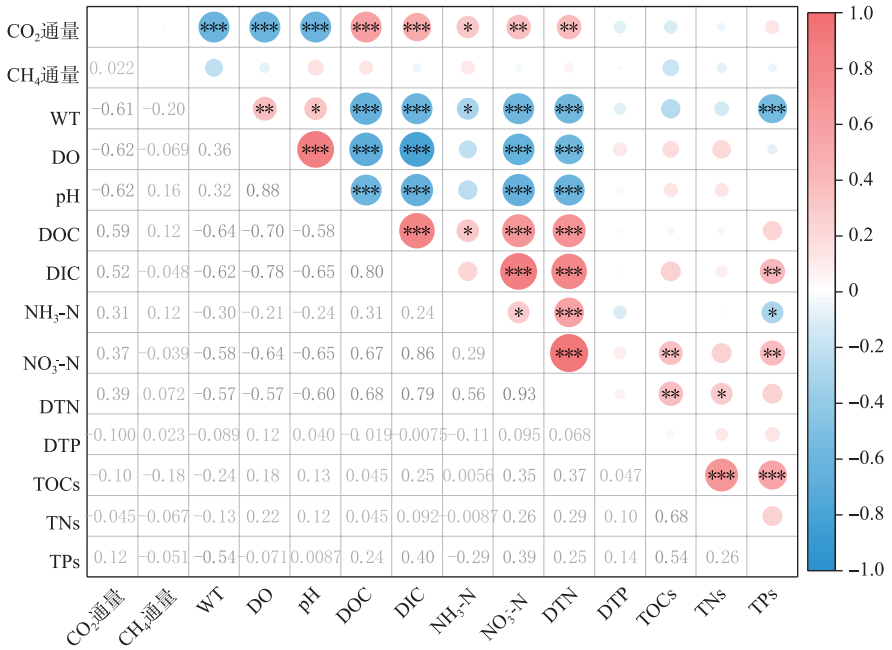

图2 CO_2、CH_4扩散通量与环境因素的相关性分析(* P<0.05, ** P<0.01, *** P<0.001)

Fig.2 Correlation analysis between CO_2, CH_4 diffusion fluxes and environmental factors

1.2 算法选择

经典机器学习中的 RF 首先被考虑,因为它在水库 CO_2通量预测方面已经表现出了较高的精确度[10,16]。SVM 是一种基于监督学习的广义线性分类器,已被广泛应用于水质评价[21]、水华预测[22]等领域,本研究首次将其用于预测水库水—气界面的温室气体通量。深度学习是指训练基于反向传播的深度神经网络(例如

全连接神经网络)的过程,其广泛应用于计算机视觉、自然语言处理和人工智能等许多领域。深度学习中最具有代表性的算法,如 CNN、循环神经网络(RNN)和 LSTM 等都是在全连接神经网络的基础上发展起来的,本研究选取 CNN 和 LSTM 作为预测算法。

RF 是一种有监督的非参数统计算法,由大量决策树分类器构成"森林",其中每个决策树分类器都是独立的,且从输入样本中采集随机向量而生成,最终以所有决策树分类器的最高排名为输出[23]。此外,特征的重要性也可以由 RF 产生。支持向量机是一种针对线性问题的判别分类器,该算法构建一个超平面,将数据分离并分类到高维空间中[24]。深度神经网络中的 CNN 得到了最广泛的研究与应用[25]。CNN 通常由卷积层和全连接层组成(图 3a),卷积层通过卷积运算提取输入特征,更深的卷积层可以从上一个被提取的特征中迭代提取更复杂的特征。每个卷积层由几个卷积单元组成,其单元参数通过反向传播算法进行优化[26]。LSTM 通过门控制与长短记忆相结合,克服了 RNN 的缺点。LSTM 的基本单元包括输入门、输出门和遗忘门(图 3b)。遗忘门检查输入 x_t 和隐藏状态 h_{t-1},并输出 0~1 之间的向量,以确定状态 S_{t-1} 中的信息是否被遗忘(式(1))。输入门中的 x_t 和隐藏状态 h_{t-1} 共同经 sigmoid 和 tanh 激活后选择候选细胞中 g_t 信息的更新(式(2)和式(3))。新细胞信息 S_t 由旧细胞信息 S_{t-1} 通过遗忘门和输入门确定(式(4))。细胞更新后,根据 x_t 和 h_{t-1} 确定细胞的输出(式(5)和式(6))。最后,CNN 和 LSTM 分别连接到全连接神经网络,将所有局部特征结合为全局特征并输出(图 3)。

$$f_t=\sigma(W_f x_t+W_f h_{t-1}+b_f) \tag{1}$$

$$i_t=\sigma(W_i x_t+W_i h_{t-1}+b_i) \tag{2}$$

$$g_t=\tanh(W_g x_t+W_g h_{t-1}+b_g) \tag{3}$$

$$S_t=g_t\otimes i_t+S_{t-1}\otimes f_t \tag{4}$$

$$o_t=\sigma(W_o x_t+W_o h_{t-1}+b_o) \tag{5}$$

$$h_t=\tanh(S_t)\otimes o_t \tag{6}$$

式中,W_f、W_i、W_g、W_o 分别为对应门的矩阵权值乘以输入 x_t 和隐藏状态 h_{t-1};b_f、b_i、b_g、b_o 为相应闸的偏置项;σ 为 sigmoid 激活函数;tan$h$ 为 tanh 激活函数。

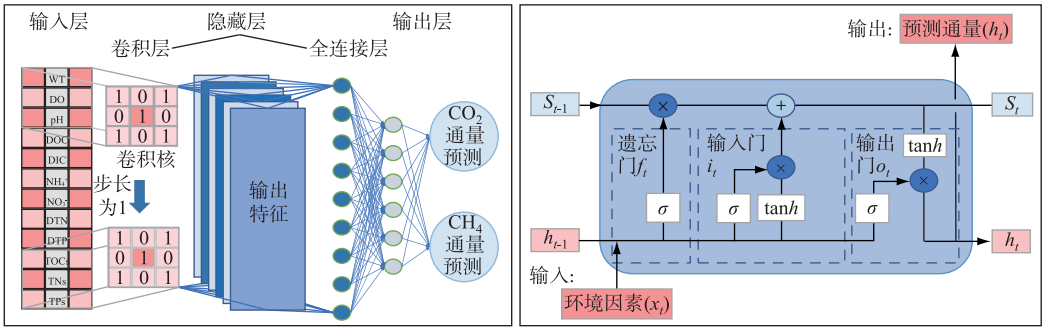

图 3 深度学习模型结构

Fig.3 Schematic diagram of deep learning model structure

1.3 预处理、模块导入、参数配置与模型训练

为了更好地拟合模型,采集的样本数据需要进行预处理。本研究原始数据来自课题组团队的前期实验数据,样本数据高度完整,避免了矩阵稀疏等问题。此外,按照 3σ 定律[27]对样本数据进行异常值检验,结果表明 CO_2 和 CH_4 扩散通量的极值都在可接受的范围内,因此无需对数据进行噪声去除。本研究对样本数据仅进行标准化处理(式(7))。

$$z^*=\frac{z-z_{\min}}{z_{\max}-z_{\min}}-0.5 \tag{7}$$

式中,z 为数据的原始值;z^* 为数据标准化后的值;$z_{\max}$ 和 $z_{\min}$ 分别为样本数据中的最大值和最小值。

本研究所有算法的源代码来自 Python 3.8 (https://www.python.org/),算法模块从基于 Tensorflow1.0 框架的 Keras 深度学习工具和 Scikit-Learn (*sklearn*)库中导入。其中,模块 RandomForestClassifier 和 SupportVectorRegression 分别从 *sklearn.ensemble* 和 *sklearn.svm* 中导入,Conv1D 和 LSTM 从 *keras.layers* 中导入。

RF 和 SVM 使用 *sklearn* 库中的默认参数配置。深度学习算法涉及特定的网络结构和激活函数选择。本研究中的 CNN 包含两个一维卷积层,每层卷积核个数为 64,移动步长为 1。两层 LSTM 每层有 64 个神经元,提取环境因子中的特征并输出 1×64 的向量。将 CNN 和 LSTM 的输出特征分别输入到 4 层全连接神经网络。全连接层的神经元节点数分别为 64、32、16 和 1,选取 tanh 函数作为激活函数。最后,反标准化后的输出就是模型的预测值。本研究采用绝对均值误差(*MAE*)作为损失函数,使用 Adam 算法优化模型中的权重和偏置,直到 *MAE* 收敛并稳定。此外,采用 RF 进行环境因子的特征重要性评估,采用决策树(DT)构建可视化的源汇分类模型。

在 4 个回归模型中,54 组样本数据被随机分为一个训练集和一个验证集。根据常用划分原则[28],训练集的样本数量为全部样本的 80%(54 组×80% ≈43 组),验证集的样本数量为 20%(剩余 11 组)。通过实际通量和预测通量之间的直接比较、相关性(R^2)及均方根误差(*RMSE*)来评估预测精度[29]。模型框架如图 4 所示。

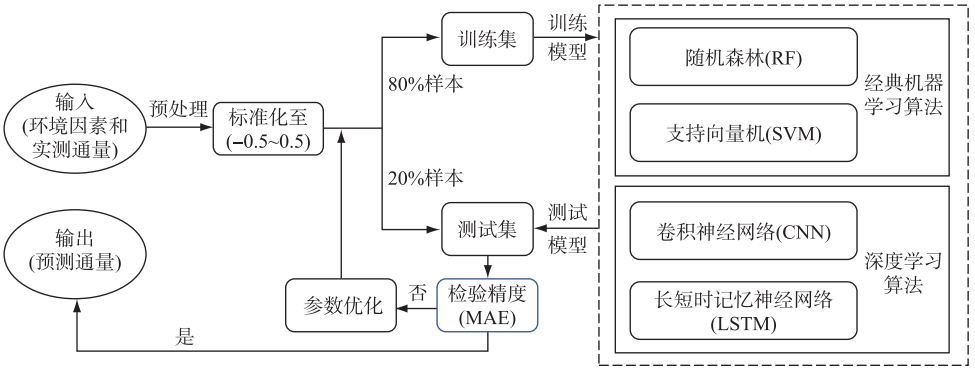

图 4 模型框架

Fig.4 The model framework of this study

1.4 k 折交叉法

k 折交叉验证法用于在有限样本数据上评估机器学习模型的稳定性和泛化性[30]。k 折交叉验证将样本数据集随机分割成 k 组,在其中随机选择 2 组作为验证集,其余 k-2 组作为训练集。k 次训练数据对应 k 个预测精度(*RMSE*),比较预测精度判断模型是否稳定。考虑到样本数据的时序特征,本研究 k 取 6。

2 结果与讨论

2.1 CO_2、CH_4扩散通量预测

为了检验不同算法预测温室气体扩散通量的可行性,采用 RF 和 SVM(经典机器学习)、CNN 和 LSTM(深度学习)对回归模型进行训练,并对其预测精度(*RMSE* 和 R^2)进行评估。4 种模型在验证集(11 个样本)中预测 CO_2扩散通量与实际 CO_2扩散通量的对比如图 5a~d 所示。值得注意的是,RF 模型的输入特征需要与算法相适应的特殊处理。因此,RF 模型与其他 3 种模型分别调试,且具有不同的验证集(由于模型已被 6 折交叉法验证过稳定性,这样的处理不会影响预测精度)。所有的验证集都包括 CO_2的源和汇。

结果表明,4 种模型均具有一定的预测精度。RF 对 CO_2源汇的预测效果最好,准确预测了通量的正负(图 5a),其余模型对 CO_2源汇的预测均有一定的失误(图 5b~d)。此外,在 CO_2扩散通量较高的样本中观察到较大的预测误差,这可能是由于在原始数据中较高的 CO_2扩散通量的出现频率较低(图 1a),从而导致较高的 CO_2扩散通量无法准确拟合。CO_2扩散通量的实测值在−0.499~9.185 mmol/(m^2·h)范围内波动,RF、

SVM、CNN 和 LSTM 4 种模型的 *RMSE* 分别为 0.452、1.204、0.570 和 0.424 mmol/(m²·h)。参考实测值的范围,本研究认为 4 种模型预测误差在可接受的范围内,预测精度(由 *RMSE* 定义)表现为 LSTM>RF>CNN>SVM(图 6a)。4 种模型的预测 CH_4 扩散通量与实际 CH_4 扩散通量的对比如图 5e~h 所示。CH_4 扩散通量较高时的预测精度较低,这与 CO_2 扩散通量的预测情况一致。RF 有最高的预测精度(*RMSE* = 0.140 μmol/(m²·h)),其次是 LSTM(0.494 μmol/(m²·h))和 CNN(0.590 μmol/(m²·h)),而 SVM 预测精度最低(0.864 μmol/(m²·h))。与 CH_4 扩散通量的实测范围相比(0.129~6.850 μmol/(m²·h)),4 种模型的预测精度也在可接受的范围内。

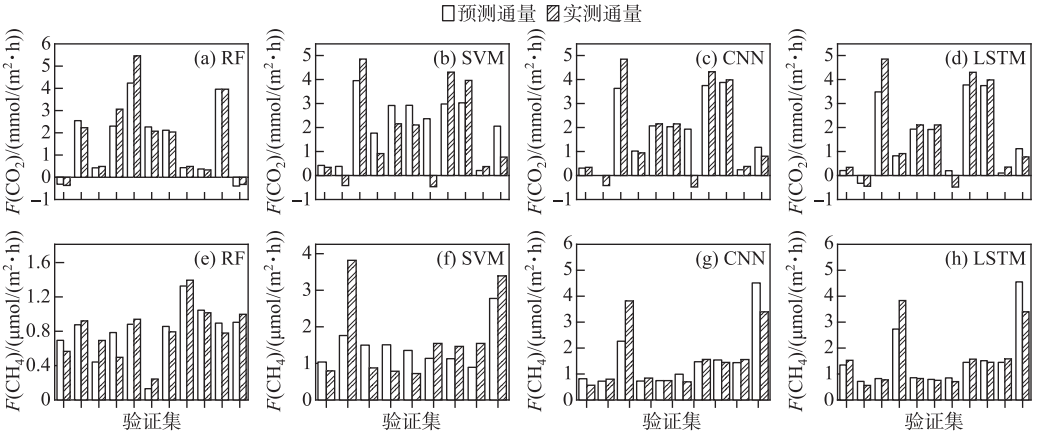

图 5 RF(a,e)、SVM(b,f)、CNN(c,g)、LSTM(d,h)预测通量与实测通量比较

Fig.5 Comparisons between the predicted flux and real flux using algorithms of RF (a and e), SVM (b and f), CNN (c and g) and LSTM (d and h)

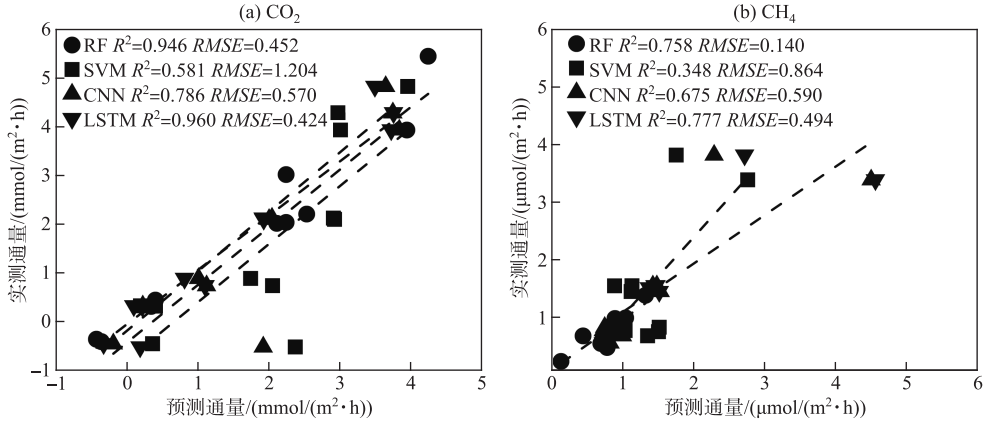

图 6 预测值与实测值相关性分析:(a) CO_2 扩散通量、(b) CH_4 扩散通量

Fig.6 Correlation analysis between predicted and real fluxes: (a) CO_2 diffusion flux, (b) CH_4 diffusion flux

4 种模型都在一定程度上证实了预测温室气体扩散通量的可行性。然而,作为前沿分支的深度学习模型的预测精度并不总是优于经典机器学习模型,特别是在 CH_4 扩散通量的预测方面,RF 取得了最好的预测效果。在过去的几年里,深度学习算法在许多领域(语言、视觉、图像处理等)的表现都远超过了经典机器学习算法[14],且已成为大多数领域的首选技术。深度学习算法具有大量的神经元和复杂的网络结构,也需要更多的输入特征和大量的数据集(数以万计),这或许可以解释本研究的结果。现有方法监测水库中的温室

气体通量需要大量的人力和时间成本,并且难以获得大量的数据集,而对于小样本数据(本研究为 756 个数据),经典机器学习模型可能会更加适用。Kia 等[31]研究表明在有限的样本数据下,经典机器学习算法预测露天煤矿地的 CH_4通量的效果更好,可以防止模型过拟合。RF 一直以来都被视为一种简单高效的经典机器学习模型,它可以处理高维数据,对数据的适应性强,其在水库[10,15-16]、农业土壤[32]、甚至厌氧消化池[33]中均有较高的预测温室气体的精度。LSTM 对 CO_2和 CH_4扩散通量的预测效果较好是因为 LSTM 内部状态细胞的存储历史信息能力[34]。SVM 是基于超平面的适用于小样本数据的经典机器学习方法,在本研究中首次被运用于水库温室气体通量的预测,但其表现出了最差的预测性能。这可能是因为 SVM 对超参数的配置和核函数选择敏感,而本研究的 SVM 模型采用 *sklearn* 库的常用默认配置,未进行超参数和核函数的调整优化。

另外,本研究发现 CO_2扩散通量的预测精度显著高于 CH_4扩散通量。相关性分析表明,CO_2扩散通量与较多环境因子显著相关而 CH_4扩散通量则相反(图 2)。水体溶解 CO_2浓度和 CO_2通量主要受初级生产、城市污水和碳酸盐电离平衡等的影响[35],这些过程均与水体理化指标的关系密切,这可能是预测精度高的重要原因,即选择了合适的输入特征。而水体 CH_4主要源自沉积物中厌氧微生物的代谢,本研究对输入特征的考虑可能不够充分,可能忽略了一些生物性因素,例如产甲烷菌或者甲烷氧化菌的丰度、群落结构和代谢等。

2.2 环境因子重要性

为了进一步探索环境因素与 CO_2和 CH_4扩散通量之间的潜在联系,本研究使用 RF 的特征重要性来评估环境因素的重要性。值得一提的是 RF 的算法原理:RF 从原始训练样本 N 中有放回地重复随机抽取 n 个样本训练决策树,并重复生成 m 棵决策树组成随机森林。特征重要性则是决策树上特征的贡献大小(基尼指数),单棵决策树对特征重要性的评估存在误差,但随机产生大量的决策树后,随机森林的输出结果基本收敛且稳定。本研究中 RF 运行 1000 次后取特征重要性的平均值,进一步确保评估结果的准确性。12 个环境因素对温室气体通量的重要性排序如图 7a~b 所示。

RF 的结果初步表明,沉积物 TOC 是影响 CO_2和 CH_4扩散通量的最重要特征,沉积物 TN 和沉积物 TP 也相对较重要。温度和 pH 值对 CO_2和 CH_4扩散通量的重要性相对较低。DO 对 CH_4扩散通量更重要而对 CO_2扩散通量则相反(图 7a~b)。然而,不难发现 12 个环境因素的重要性差异较小,这是因为在使用 RF 评估重要性时,需要考虑输入特征之间的相关性,并适当选择输入特征[36]。程麒铭等[37]使用 RF 算法量化了生物滞留系统中环境因子的重要性,并表明 TN 与 NH_3-N、NO_3^--N 之间的强相关性降低了它们的重要性。在本研究中,考虑到环境因素的相关性和代表性,选取合适的环境因素作为 RF 的输入特征。WT 与大多数环境因素显著相关(P<0.05)(图 2),但 WT 是研究区气候的直接代表,故选为气候因子。DO 与 pH 呈极显著正相关(P<0.001)。考虑到 DO 对 CH_4通量的高度重要性(图 7b),选择 DO 作为水环境因子。不同形态的 C、N、P 之间存在显著的强相关性,最终选择 DTN 作为营养因子。沉积物中 TOC 与沉积物中 TN 和 TP 的相关性显著(P<0.001),因此选择 TOC 作为沉积物因子。将上述筛选后的 4 类环境因子再次输入 RF 进行重要性评估,结果如图 7c~d 所示。结果表明,沉积物因子和营养物质因子对 CO_2和 CH_4扩散通量的影响都更重要,而水环境因子和气候因子的影响相对较小。显然,选择一个环境因子来代表一个类别,可以大大提高输出结果的可靠性和可分析性。例如,输入全部环境因素会使营养因子被划分为 6 个自相关性显著的环境因素(DOC、DIC、NH_3-N、NO_3^--N、DTN 和 DTP),输出结果似乎平均地分割了营养物质的重要性(图 7a 和 7b)。

沉积物中微生物的好氧和厌氧代谢分别是水生生态系统中 CO_2和 CH_4的重要来源[38-39],尤其是 CH_4。CH_4由厌氧沉积物中的产甲烷菌产生,并通过植物运输、气泡排放和扩散等方式进入大气[40],且一般认为 CH_4通量与沉积物 TOC 呈正相关[41],这与本研究的结论一致,即沉积物对温室气体通量的最高重要性。水中的氮素同样会促进温室气体的产生与排放[36,42]。此外,DIC 包含 CO_3^{2-}、HCO_3^-和溶存 CO_2等无机碳,而 DOC 为水中微生物的代谢提供有机碳源,这两者也都是水中 CO_2的主要来源。气候因子和水环境因子通常也是影响细菌代谢和温室气体通量的重要因素[43-44],本研究中相关性分析也表明水温、DO 和 pH 与 CO_2扩散通量呈极显著相关(P<0.001)(图 2)。然而,RF 的输出结果表明它们的重要性相较低于沉积物因子和营养因子。

Qin 等[36]研究表明 RF 将 NO_3^--N 定义为对 CH_4扩散通量重要性最高的因素,这与本研究结论类似。此外,Hyungseok 等[10]研究表明,仅输入 3 个特征(电导率、DO 和 TOC)的 RF 模型解释了 84.4% 的 CO_2通量时间变化,且电导率被 RF 确定为影响 CO_2通量的关键因素。Mosher 等[16]使用 RF 模型得出,表层水 pH、底层水 DOC、距大坝的距离和表层水 DOC 是 CO_2通量的最重要的影响因子,而底层水温度、表层水 O_2、表层水 pH 和表层水 DOC 是影响 CH_4通量的重要因素。但 Mosher 等[16]也指出,其输入特征只考虑了水体理化指标,适当增加输入特征(例如沉积物)可以改善模型预测和评估性能,这一点在本研究中被证实。

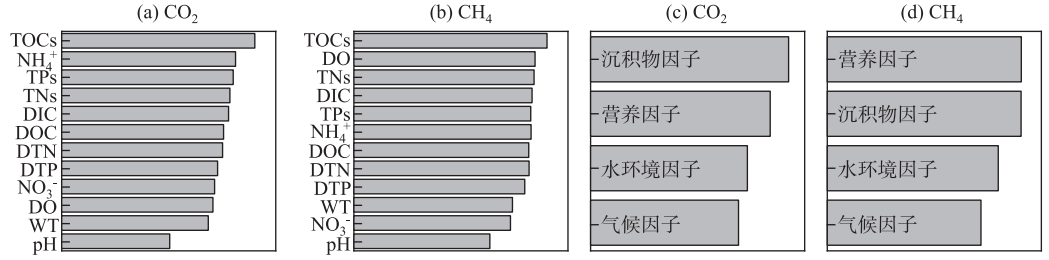

图 7 环境因子对 CO_2、CH_4扩散通量的重要性

Fig.7 The importance of environmental factors to CO_2, CH_4 fluxes

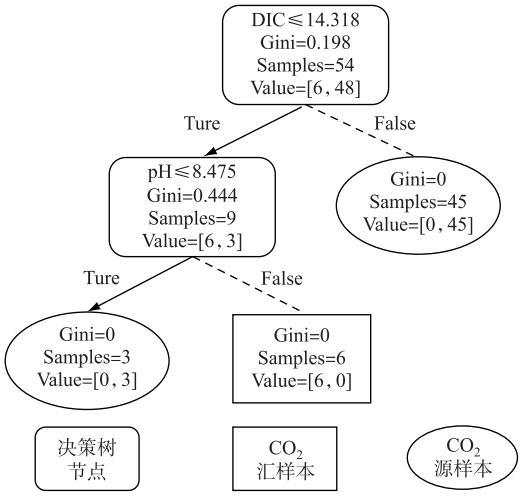

图 8 决策树结果

Fig.8 The results of the decision tree

2.3 决策树模型

本研究原始样本可以分类为 CO_2扩散通量的源和汇,但是均为 CH_4的源(图 1),因此,采用所有环境因素和 CO_2扩散通量构建决策树分类模型,从而可视化地预测 CO_2扩散通量的源汇。决策树结果如图 8 所示,树最大深度为 2,本研究选择基尼指数作为划分标准。决策树对 54 个样本均进行了正确分类,其中 CO_2汇样本 6 个,源样本 48 个,分类正确率高达 100%。决策树首先选择 DIC 为特征对所有样本进行分割,将 DIC 大于 14.318 mg/L 的 45 个样品直接归类为 CO_2源。之后 pH 以 8.475 为阈值将左子树划分为相对酸性或碱性条件,剩余的 9 个样本根据 pH 划分为 6 个 CO_2汇样本和 3 个 CO_2源样本。最后,决策树的所有叶节点都是基尼指数为 0 的纯节点。虽然该决策树仅使用两个环境因子(DIC 和 pH)对 CO_2的源汇进行分类,但其准确率高达 100%,这表明决策树是一种简便、可靠并且可视化的 CO_2通量分类方法。

决策树结果表明,低浓度 DIC 和相对碱性条件有利于水体吸收大气 CO_2且成为 CO_2的汇。DIC 包含 CO_3^{2-}、HCO_3^-、H_2CO_3和溶存 CO_2等无机碳,且构成水中无机碳的电离平衡($CO_2 + H_2O \rightleftharpoons H_2CO_3 \rightleftharpoons H^+ + HCO_3^- \rightleftharpoons 2H^+ + CO_3^{2-}$)。本研究中,当 DIC 超过 14.318 mg/L 时,水体 CO_2浓度过饱和迫使水体向大气释放 CO_2并成为其源,而当 DIC 小于 14.318 mg/L 时,CO_2的源汇由 pH 值决定。酸性条件有利于碳酸盐平衡向左偏移,从而促进水中 CO_2浓度的过饱和,并释放成为其源。

3 结论与展望

3.1 结论

1)LSTM 和 RF 分别产生了最优的 CO_2和 CH_4扩散通量预测精度,均方根误差(*RMSE*)分别为 0.424

mmol/(m^2·h)和 0.140 μmol/(m^2·h),预测值与实测值的 R^2分别为 0.960 和 0.758(P<0.001)。经典机器学习算法中 RF 预测效果显著地优于 SVM。

2) RF 的特征重要性评估表明沉积物因子和营养因子均为影响 CO_2和 CH_4扩散通量的关键因子,气候因子和水环境因子相较次之。

3)决策树描绘了决定 CO_2扩散通量源和汇的环境因子的极限阈值,对所有样本的分类准确性高达 100%。低浓度的 DIC 和碱性条件有利于水体成为 CO_2的汇。

4)本研究将先进的人工智能算法引入了水库温室气体排放的研究和实践中,证实其应用于预测水库温室气体通量的可靠性和准确性。随着计算机和人工智能的发展,使用机器学习解决实际问题必将成为该领域未来的研究热点。

3.2 不足与展望

本研究提出了采用几种机器学习算法(尤其是深度学习)预测水库水—气界面温室气体扩散通量,迈出了该领域向人工智能结合的重要一步。由于计算机和人工智能的大力发展以及水库温室气体通量监测的较大难度,本研究认为在未来的几年里,基于机器学习的水库温室气体排放预测必将成为一个新的研究热点。但是本研究仍存在以下不足:(i)本研究的实测通量采用薄边界模型计算所得,该模型对表面扩散速率和通量的估算精度相对低于静态箱法[45],也可能因为对风速和流速等把握不充分而使估算的扩散通量存在误差。(ii)由于温室气体通量监测的较大难度和野外采样的不确定性,本研究的原始数据仅有 54 个样本(756 个数据),属于小样本数据下的模型构建,这在一定程度会限制模型的预测精度。(iii)尽管本研究的输入特征考虑了水和沉积物中的重要物理化学指标,但是一些未被考虑的因素可能会提升模型的预测性能。例如水库的流态、库龄、富营养化程度等。(iv)模型构建的部分超参数选用了 python 中 *sklearn* 库的默认参数,为了公平地对比不同算法的预测精度,本研究没有过多地调整超参数来优化具体的算法。在未来的研究中,团队会加大监测频率和采样点数量,考虑更多的输入特征,专注于参数优化下算法预测性能的提升,希望能提高水库温室气体排放的预测模型的准确性。

尽管有上述不足,但是本研究证实了机器学习(尤其是深度学习)用于预测水库温室气体排放的可靠性。首先,机器学习模型将从全新的角度增进对环境因素和温室气体通量的因果关系的理解,例如采用随机森林算法估算各类因素的重要性,采用决策树对温室气体的源和汇进行基于极限阈值的可视化分类等。其次,相较于针对少数采样点的传统监测,机器学习模型的预测可以使得温室气体排放估算在空间和时间上连续化,这会增加一定年限内对水库温室气体排放估算的精确度,甚至可以精准估算未采样地区的排放通量。最后,机器学习模型从一定程度上可以预测一些环境变化下的排放量变化,例如全球变暖、富营养化加剧、水库老化和运行等。精准的机器学习模型甚至有助于水库的全生命周期分析,估算水利工程完整的碳足迹。

4 参考文献

[1] Guérin F, Abril G, Richard S *et al*. Methane and carbon dioxide emissions from tropical reservoirs: Significance of downstream rivers. *Geophysical Research Letters*, 2006, **33**(21): L21407. DOI: 10.1029/2006GL027929.

[2] Tranvik LJ, Downing JA, Cotner JB *et al*. Lakes and reservoirs as regulators of carbon cycling and climate. *Limnology and Oceanography*, 2009, **54**(6part2): 2298-2314. DOI: 10.4319/lo.2009.54.6_part_2.2298.

[3] Sawakuchi HO, Bastviken D, Sawakuchi AO *et al*. Oxidative mitigation of aquatic methane emissions in large Amazonian rivers. *Global Change Biology*, 2016, **22**(3): 1075-1085. DOI: 10.1111/gcb.13169.

[4] Deemer BR, Harrison JA, Li SY *et al*. Greenhouse gas emissions from reservoir water surfaces: A new global synthesis. *BioScience*, 2016, **66**(11): 949-964. DOI: 10.1093/biosci/biw117.

[5] Ssemiganda M. Methane and nitrous oxide emissions from subtropical coastal systems and freshwater reservoirs[Dissertation]. University of Queensland Library, 2013. DOI: 10.14264/uql.2014.157

[6] Prairie YT, Alm J, Beaulieu J *et al*. Greenhouse gas emissions from freshwater reservoirs: What does the atmosphere see? *Ecosystems: New York*, 2018, **21**(5): 1058-1071. DOI: 10.1007/s10021-017-0198-9.

[7] Cai WJ, Li K, Liao H *et al*. Weather conditions conducive to Beijing severe haze more frequent under climate change. *Nature Climate Change*, 2017, **7**(4): 257-262. DOI: 10.1038/nclimate3249.

[8] Raymond PA, Caraco NF, Cole JJ. Carbon dioxide concentration and atmospheric flux in the Hudson River. *Estuaries*, 1997, **20**(2): 381. DOI: 10.2307/1352351.

[9] Chung S, Park HS, Yoo JS *et al*. Variability of pCO_2 in surface waters and development of prediction model. *Science of the Total Environment*, 2018, **622/623**: 1109-1117. DOI: 10.1016/j.scitotenv.2017.12.066.

[10] Hyungseok P, Sewoong C, Sungjin K *et al*. Effect of buoyant turbulence and water quality factors on the CO_2 net atmospheric flux changes in a stratified reservoir. *Science of the Total Environment*, 2021, **776**: 145940. DOI: 10.1016/j.scitotenv.2021.145940.

[11] DelSontro T, Beaulieu JJ, Downing JA. Greenhouse gas emissions from lakes and impoundments: Upscaling in the face of global change. *Limnology and Oceanography Letters*, 2019, **3**(3): 64-75. DOI: 10.1002/lol2.10073.

[12] Krogh A. What are artificial neural networks? *Nature Biotechnology*, 2008, **26**(2): 195-197. DOI: 10.1038/nbt1386.

[13] Portugal I, Alencar P, Cowan D *et al*. The use of machine learning algorithms in recommender systems: A systematic review. *Expert Systems with Applications*, 2018, **97**: 205-227. DOI: 10.1016/j.eswa.2017.12.020.

[14] Halbouni A, Gunawan TS, Habaebi MH *et al*. Machine learning and deep learning approaches for CyberSecurity: A review. *IEEE Access*, **10**: 19572-19585. DOI: 10.1109/ACCESS.2022.3151248.

[15] Chen ZH, Ye XQ, Huang P. Estimating carbon dioxide (CO_2) emissions from reservoirs using artificial neural networks. *Water*, 2018, **10**(1): 26. DOI: 10.3390/w10010026.

[16] Mosher J, Fortner A, Phillips J *et al*. Spatial and temporal correlates of greenhouse gas diffusion from a hydropower reservoir in the southern United States. *Water*, 2015, **7**(11): 5910-5927. DOI: 10.3390/w7115910.

[17] Li Z, Lu LH, Lv PY *et al*. Imbalanced stoichiometric reservoir sedimentation regulates methane accumulation in China's Three Gorges Reservoir. *Water Resources Research*, 2020, **56**(9): e2019WR026447. DOI: 10.1029/2019WR026447.

[18] Sun HY, Yu RH, Liu XY *et al*. Drivers of spatial and seasonal variations of CO_2 and CH_4 fluxes at the sediment water interface in a shallow eutrophic lake. *Water Research*, 2022, **222**: 118916. DOI: 10.1016/j.watres.2022.118916.

[19] Liu J, Xiao SB, Wang CH *et al*. Spatial and temporal variability of dissolved methane concentrations and diffusive emissions in the Three Gorges Reservoir. *Water Research*, 2021, **207**: 117788. DOI: 10.1016/j.watres.2021.117788.

[20] 王雨潇. 三峡库区万州段干流、典型支流 CO_2、CH_4通量变化特征研究[学位论文]. 重庆: 重庆交通大学, 2020.

[21] Li W, Yang MY, Liang ZW *et al*. Assessment for surface water quality in Lake Taihu Tiaoxi River Basin China based on support vector machine. *Stochastic Environmental Research and Risk Assessment*, 2013, **27**(8): 1861-1870. DOI: 10.1007/s00477-013-0720-3.

[22] Alúez FMB, Palenzuela JMT, Spyrakos E *et al*. Machine learning methods applied to the prediction of pseudo-*Nitzschia* spp. blooms in the Galician rias baixas (NW Spain). *ISPRS International Journal of Geo-Information*, 2021, **10**(4): 199. DOI: 10.3390/ijgi10040199.

[23] Che DS, Liu Q, Rasheed K *et al*. Decision tree and ensemble learning algorithms with their applications in bioinformatics. *Advances in Experimental Medicine and Biology*, 2011, **696**: 191-199. DOI: 10.1007/978-1-4419-7046-6_19.

[24] Ding SF, Zhu ZB, Zhang XK. An overview on semi-supervised support vector machine. *Neural Computing and Applications*, 2017, **28**(5): 969-978. DOI: 10.1007/s00521-015-2113-7.

[25] Gu JX, Wang ZH, Kuen J *et al*. Recent advances in convolutional neural networks. *Pattern Recognition*, 2018, **77**: 354-377. DOI: 10.1016/j.patcog.2017.10.013.

[26] Tian CJ, Ma J, Zhang CH *et al*. A deep neural network model for short-term load forecast based on long short-term memory network and convolutional neural network. *Energies*, 2018, **11**(12): 3493. DOI: 10.3390/en11123493.

[27] Ma J, DingYX, Cheng JCP *et al*. Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques. *Water Research*, 2020, **170**: 115350. DOI: 10.1016/j.watres.2019.115350.

[28] Qin Y, Ouyang CY, Fang P. Reservoir carbon dioxide flux prediction based on CNN-LSTM model and small sample datas. *Journal of Chongqing Jiaotong University: Natural Science*, 2022, **41**(6): 119-125. [秦宇, 欧阳常悦, 方鹏. 基于 CNN-LSTM 模型及小样本数据的水库二氧化碳通量预测. 重庆交通大学学报: 自然科学版, 2022, **41**(6): 119-125.]

[29] Barnston AG. Correspondence among the correlation, RMSE, and heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting*, 1992, **7**(4): 699-709. DOI: 10.1175/1520-0434(1992)0070699: catcra>2.0.co;2.

[30] Lalwani P, Mishra MK, Chadha JS *et al*. Customer churn prediction system: A machine learning approach. *Computing*, 2022, **104**(2): 271-294. DOI: 10.1007/s00607-021-00908-y.

[31] Kia S, Nambiar MK, Thé J *et al*. Machine learning to predict area fugitive emission fluxes of GHGs from open-pit mines. *Atmosphere*, 2022, **13**(2): 210. DOI: 10.3390/atmos13020210.

[32] Hamrani A, Akbarzadeh A, Madramootoo CA *et al*. Machine learning for predicting greenhouse gas emissions from agricultural soils. *Science of the Total Environment*, 2020, **741**: 140338. DOI: 10.1016/j.scitotenv.2020.140338.

[33] Wang LG, Long F, Liao W *et al*. Prediction of anaerobic digestion performance and identification of critical operational parameters using machine learning algorithms. *Bioresource Technology*, 2020, **298**: 122495. DOI: 10.1016/j.biortech.2019.122495.

[34] Yang BW, Xiao ZJ, Meng QJ *et al*. Deep learning-based prediction of effluent quality of a constructed wetland. *Environmental Science and*

Ecotechnology, 2023, **13**: 100207. DOI: 10.1016/j.ese.2022.100207.

[35] Zhao ZQ, Mo YS, Jiao SL *et al*. Analysis of source and change of dissolved inorganic carbon in Guangzhao reservoir. *Research of Environmental Sciences*, 2020, **33**(12): 2810-2819. DOI: 10.13198/j.issn.1001-6929.2020.08.10. [赵宗权, 莫跃爽, 焦树林等. 光照水库溶解无机碳变化及其来源解析. 环境科学研究, 2020, **33**(12): 2810-2819.]

[36] Qin Y, Ouyang CY, Gou YJ *et al*. The characteristics and influencing factors of dissolved methane concentrations in Chongqing's central urban area in the Three Gorges Reservoir, China. *Environmental Science and Pollution Research International*, 2022, **29**(47): 72045-72057. DOI: 10.1007/s11356-022-20822-w.

[37] Cheng QM, Chen Y, Liu Z *et al*. Multi-objective evaluation method of bioretention system based on random forest-projection pursuit method. *Journal of Water Resources and Water Engineering*, 2022, **33**(4): 85-90, 96. [程麒铭, 陈垚, 刘臻等. 基于随机森林-投影寻踪法的生物滞留系统多目标评价方法. 水资源与水工程学报, 2022, **33**(4): 85-90, 96.]

[38] Bastviken D, Ejlertsson J, Tranvik L. Measurement of methane oxidation in lakes: A comparison of methods. *Environmental Science & Technology*, 2002, **36**(15): 3354-3361. DOI: 10.1021/es010311p.

[39] Oswald K, Milucka J, Brand A *et al*. Light-dependent aerobic methane oxidation reduces methane emissions from seasonally stratified lakes. *PLoS One*, 2015, **10**(7): e0132574. DOI: 10.1371/journal.pone.0132574.

[40] Zhang LH, Song CC, Wang DX. Effects of nitrogen fertilization on carbon balance in the freshwater marshes. Environmental. Science, 2006, **27**(7): 1257-1263.

[41] Li LL, Xue B, Yao SC. The significance and application of the research on production and oxidation of mathane in lake sediments. *Bulletin of Mineralogy, Petrology and Geochemistry*, 2016, **35**(4):634-645.

[42] Semrau JD, DiSpirito AA, Yoon S. Methanotrophs and copper. *FEMS Microbiology Reviews*, 2010, **34**(4): 496-531. DOI: 10.1111/j.1574-6976.2010.00212.x.

[43] Dinsmore KJ, Billett MF, Dyson KE. Temperature and precipitation drive temporal variability in aquatic carbon and GHG concentrations and fluxes in a peatland catchment. *Global Change Biology*, 2013, **19**(7): 2133-2148. DOI: 10.1111/gcb.12209.

[44] Treat CC, Wollheim WM, Varner RK *et al*. Temperature and peat type control CO_2 and CH_4 production in Alaskan permafrost peats. *Global Change Biology*, 2014, **20**(8): 2674-2686. DOI: 10.1111/gcb.12572.

[45] Duchemin E, Lucotte M, Canuel R. Comparison of static chamber and thin boundary layer equation methods for measuring greenhouse gas emissions from large water bodies. *Environmental Science & Technology*, 1999, **33**(2): 350-357. DOI: 10.1021/es9800840.