DOI: 10.13671/j.hjkxxb.2022.0125
任树顺,高萌,王煦雯, 等.2022.基于3种时间序列模型的九龙江河流库区藻华预测[J].环境科学学报,42(11):172-183
REN Shushun,GAO Meng,WANG Xuwen,*et al*.2022.Algal bloom prediction in the Jiulong River Reservoir based on three types of time series models [J].Acta Scientiae Circumstantiae,42(11):172-183

基于3种时间序列模型的九龙江河流库区藻华预测

任树顺[1,2],高萌[3],王煦雯[4],余镒琦[1,2],陈纪新[2],陈能汪[1,2,*]

1. 厦门大学环境与生态学院,福建省海陆界面生态环境重点实验室,厦门 361102
2. 厦门大学,近海海洋环境科学国家重点实验室,厦门 361102
3. 罗德岛大学,海洋学研究生院,纳拉甘西特,美国罗德岛 02882
4. 中国海洋大学环境科学与工程学院,青岛 266100

摘要:湖库富营养化和有害藻华是全球性生态环境问题,藻华预测与早期预警是保障湖库水源地供水安全的关键技术.如何基于高频水生态在线监测数据进行藻华的实时动态预测成为水生态管理领域的重大需求.本研究以福建省九龙江江东库区(水源地)为例,利用3年连续观测的逐时平均总叶绿素a浓度数据,对比研究了SARIMA、Prophet和LSTM(长短期记忆神经网络)3种时间序列模型在藻华(日平均叶绿素a大于15 μg·L^{-1})预测方面的效果.结果表明:①时间序列模型要求参数少,灵活性强,能清晰反映水质特征和未来变化趋势,可弥补传统藻类监测预警方法的局限性;②基于深度学习框架的LSTM模型,具有独特的迭代优化算法,对藻类非线性变化特征的识别和预测能力较强,其总叶绿素a逐日预测和7日预测效果均显著优于SARIMA模型和Prophet模型;③输入数据长度会在一定程度上影响模型预测效果,最优的输入数据时间长度为7 d;输入数据频率对预测效果也有影响,在预测非藻华日时,小时数据的预测效果优于日频率数据;在预测藻华日时,两种频率数据无显著差异,但日频率数据能更准确识别藻华日特征.总结起来,基于LSTM模型实现总叶绿素a浓度的短期预测,可为九龙江河流库区藻华早期预警和供水安全保障提供技术支撑.

关键词:藻华;预测模型;人工神经网络;高频监测;叶绿素a;九龙江

文章编号:0253-2468(2022)11-0172-12 **中图分类号**:X524,X32 **文献标识码**:A

Algal bloom prediction in the Jiulong River Reservoir based on three types of time series models

REN Shushun[1,2], GAO Meng[3], WANG Xuwen[4], YU Yiqi[1,2], CHEN Jixin[2], CHEN Nengwang[1,2,*]

1. Fujian Provincial Key Laboratory for Coastal Ecology and Environmental Studies, College of the Environment and Ecology, Xiamen University, Xiamen 361102
2. State Key Laboratory of Marine Environmental Science, Xiamen University, Xiamen 361102
3. Graduate School of Oceanography, University of Rhode Island, Narragansett, RI 02882
4. College of Environmental Science and Engineering, Ocean University of China, Qingdao 266100

Abstract: Eutrophication and harmful algal blooms in lakes and reservoirs are global eco-environmental issues. The prediction and early warning of algal blooms are the key techniques for securing the safe drinking water supply. How to predict algal blooms in a real-time dynamic way based on high-frequency water ecology monitoring data has become a major demand in the field of aquatic ecosystem management. Taking Jiangdong reservoir of Jiulong River (i.e., drinking water source of Xiamen in Fujian Province) as a case study, this study developed and compared the performance of three types of time series models of SARIMA, Prophet, and LSTM (long-term and short-term memory neural network) in predicting algal bloom (defined as daily average chlorophyll-a is greater than 15 μg·L^{-1}), using the three-year continuously observed hourly mean total chlorophyll-a concentration data. The results show that: ①the time series model requires few parameters and has strong flexibility, which reflect the water quality characteristics and future trends, and can overcome the limitations of traditional methods of algae monitoring and early warning; ②The LSTM model based on the deep learning framework has a relatively strong ability to identify and predict the nonlinear variation characteristics of algae, due to its unique iterative optimization algorithm; the LSTM performance on daily prediction and seven-day prediction of total chlorophyll-a are both better than SARIMA model and Prophet model; ③The length of input data will affect the prediction performance of the models to some extent. The optimal length of inputs in this

收稿日期:2022-02-15 **修回日期**:2022-04-11 **录用日期**:2022-04-12
基金项目:国家自然科学基金(No.51961125203);中央高校基本科研业务费(No.20720200121);福建省环保科技项目(No.2021R009)
作者简介:任树顺(2000—),男, E-mail:33120182202123@stu.xmu.edu.cn; *责任作者,E-mail:nwchen@xmu.edu.cn

study was identified as 7-days. The frequency of input data also has an impact on the prediction performance. When predicting non-algal bloom days, the prediction ability to use hourly data is better than that of using daily data. When predicting algal bloom days, there is no significant difference between the two-frequency data, but the daily data can more accurately capture the characteristics of algal bloom. In summary, the short-term prediction of total chlorophyll-a concentration based on the LSTM model can provide technical support for early warning of algal bloom and water supply security in the Jiulong River Reservoir.

Keywords: algal bloom; prediction model; artificial neural network; high-frequency monitoring; chlorophyll-a; Jiulong River

1　引言(Introduction)

湖库富营养化和有害藻华(HABs)是全球性生态环境问题.有害藻华发生时,可导致水生态系统食物链断裂,大量植物、水生生物衰亡甚至灭绝(于洋等,2017),同时可能产生微囊藻毒素威胁饮用水供水安全(Ly *et al.*,2021).近年来有害藻华在我国淡水湖泊和河流中发生的频率有所增加(陈能汪,2010),有害藻华对生态系统和公共健康的影响正在加剧(Griffith *et al.*,2020).藻华预测预警和应急处置能力不足极可能加重水质污染,造成人体健康危害和巨大的经济损失(Zohdi *et al.*,2019).湖库藻华预测作为藻华预警的前置环节,是实现水质智能决策与管理的重要手段.在大数据时代,研发基于机器学习的藻华预测模型并促进业务化应用,对于构建现代化的流域水生态监管体系、保障供水安全具有十分重要的现实意义.

传统的藻华预测技术在应用方面存在局限性.以生态动力学模型(Kim *et al.*,2017;Zohdi *et al.*,2019)、藻细胞生长模型(Atzori *et al.*,2021)为基础的藻华预测技术,由于藻类复杂的生长机制和群落动态(Ma *et al.*,2013),往往需要大量的物理、化学、生物实测数据来率定模型,这类数据获取难,且模型运行要求配备专业性人员,从而大大限制了业务化应用.近年来,随着在线监测技术和计算机水平的发展,基于时间序列数据的模型开始应用于生态环境领域,这类模型能动态反映水质特征及未来变化趋势(陈能汪等,2021).基于藻类时间序列数据建模成本较低,操作方便(Zohdi *et al.*,2019).通过对历史监测数据特征的学习,时间序列模型可识别库区水环境和水华爆发规律的复杂性(Tian *et al.*,2019;Shin *et al.*,2020),对藻类未来变化趋势进行有效预测(Hochreiter *et al.*,1997).研发时间序列模型并应用于水生态管理,是解决湖库生态环境问题的努力方向(Lee *et al.*,2018).SARIMA (Seasonal Autoregressive Integrated Moving Average,季节性差分整合移动平均自回归模型)是经典的时间序列预测模型,能较好地识别数据的季节性(Rabbani *et al.*,2021),适合随机性较强的时间序列,在时间序列参数化预测方法(Parametric methods)中表现最佳(Parmezan *et al.*,2019);Prophet模型采用了Facebook数据科学团队研发的时间序列预测算法,对藻类数据的异常值和大幅度变化具有较高的包容性(Taylor *et al.*,2017;Aditya *et al.*,2021),自动化的参数设置使得训练时间更短(Toharudin *et al.*,2021).随着人工智能和机器学习算法的发展,深度神经网络LSTM(Long Short-term Memory,长短期记忆神经网络)较好地解决了循环神经网络(Recurrent Neural Network,RNN)中梯度消失和梯度爆炸的问题,有效提高了时间序列数据的预测精度,近年来在众多学科领域得到广泛应用(Huang *et al.*,2019;Yang *et al.*, 2019;Bouktif *et al.*,2020;Sangiorgio *et al.*,2020;汪凯翔等,2020;Yang *et al.*,2021).目前,基于时间序列模型的藻华预测研究主要集中在湖泊(杨昆等,2016;Wang *et al.*,2017;Daghighi,2017;Wang *et al.*,2019; Huang *et al.*, 2020)、河流(Lee *et al.*,2018;Shin *et al.*,2020;Ly *et al.*,2021)、近海(Rostam *et al.*,2021)及室内环境(Saboe *et al.*, 2021),河流库区的应用案例不多,且SARIMA、Prophet和LSTM 3个模型同时在河流库区的应用尚未见报道.

九龙江是福建省第二大河流,由北溪、西溪和南溪3条主要河流汇合而成,其中,北溪的江东库区引水工程为厦门市提供了80%左右的水源(Chen *et al.*,2021).近30年来,流域内梯级电站开发、工农业生产和生活污染排放加剧了河流库区富营养化.2009年北溪发生了大范围的拟多甲藻水华事件,严重威胁到供水安全(Li *et al.*,2011).近年来,江东库区藻华时有发生,水生态灾害风险大,是跨区域供水安全的主要隐患.由于叶绿素a是衡量浮游植物丰度或藻类生物量的重要指标,藻华可以由水中叶绿素a浓度进行表征和预测(Park *et al.*,2015).

因此,本文基于九龙江江东库区3年的高频水生态监测数据(总叶绿素a),开展藻华预测模型研究.通过构建SARIMA 、Prophet和LSTM 3种时间序列模型,对比分析其预测效果,重点考察基于深度学习框架的

LSTM 模型在叶绿素 a 预测性能方面的优势和影响因素.以期为九龙江或相似河流库区的藻华早期预警和水生态管理提供技术支撑.

2 材料与方法(Materials and methods)

2.1 数据来源与预处理

利用厦门大学自行研制的水生态在线监测系统(AquaSOO)开展连续监测,获取分钟级的水质多参数和总叶绿素 a 数据.监测系统安装在漳州市九龙江北溪江东库区(厦门市水质自动站),该库区属于狭长河道型库区(图 1).藻类监测的工作模式为流通式连续检测,测量频率为每分钟一次,泵水系统采水口位于水下 0.5~1 m 处,水样经除泡装置后进入集成多种传感器的黑色避光流路中.该系统每天定时使用清洗液对传感器和管路进行清洗,抑制生物附着.仪器每 2 周人工维护一次,用水质分析仪(Multi3410,WTW,德国)和室内仪器的测定结果进行比对和校正.本文采用 2017 年 7 月—2020 年 6 月共 3 年的总叶绿素 a 浓度数据(单位为 $\mu g \cdot L^{-1}$)进行模型研究.

原始监测数据预处理包括 3 个步骤:异常值处理、缺失值处理及数据频率转换.异常值处理流程如下:使用无效值检验去除非数值型数据后,运用界限值和分位数检验筛查出离群值,并结合人工审核判断异常值.针对单点及小段数据缺失,使用线性插值法进行插补.为保证数据频率的一致性,将分钟级数据重采样,转换为小时频率和日频率两种数据供模型研究使用.

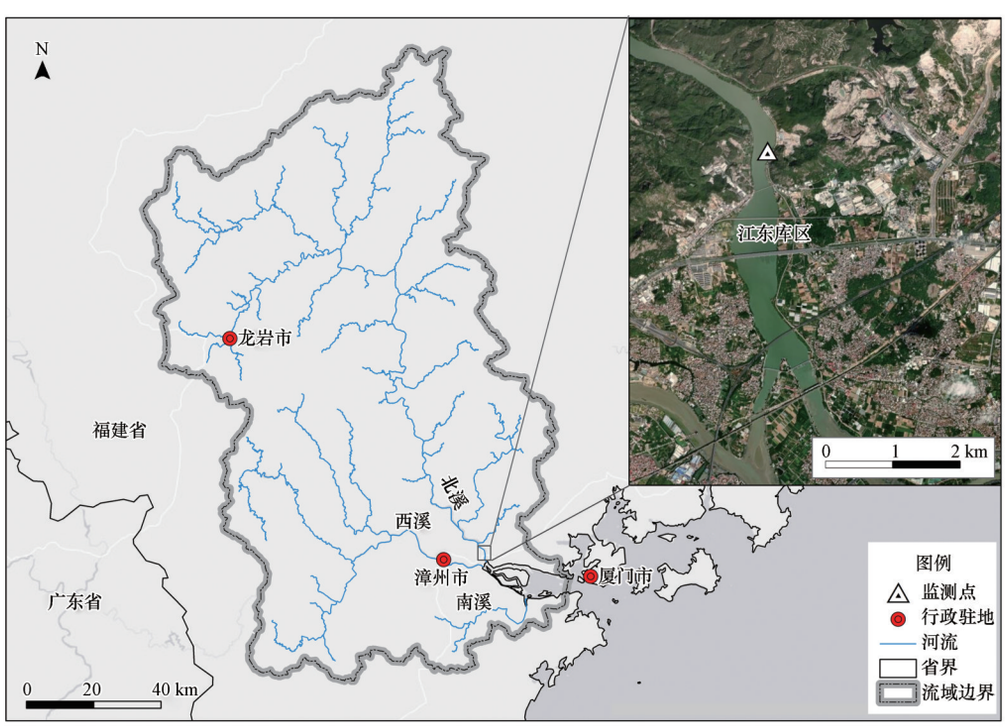

图 1 九龙江江东库区水质自动监测站地理位置图

Fig.1 Location of the automatic water quality monitoring station in the Jiangdong Reservoir of the Jiulong River

2.2 模型原理与建模流程

2.2.1 SARIMA 模型 SARIMA 模型是一种常用的时间序列预测方法,模型表达式见式(1).

$$\Phi_p(L)A_P(L^s)\left(\Delta^d\Delta_s{}^D y_t\right)=\Theta_q(L)B_Q(L^s)u_t \tag{1}$$

式中,$\Phi_p(L)$和$A_P(L^s)$分别为非季节与季节自回归算子或自回归特征多项式,其展开式分别见式(2)、式(3);$\Theta_q(L)$和$B_Q(L^s)$分别为非季节与季节移动平均算子或移动平均特征多项式,其展开式分别见式(4)、式(5);Δ、Δ_s分别为非季节和以 s 为周期的季节性差分;d、D 分别为非季节和季节性差分次数,用于将 y_t 转换为平稳

的时间序列;$u_t \sim i.i.d(0,\sigma^2)$是白噪声;下标$p$、$P$、$q$、$Q$分别为非季节、季节、自回归及移动平均算子的最大滞后阶数.

$$\Phi_p(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p \tag{2}$$

$$A_P(L^s) = 1 - \alpha_1 L^s - \alpha_2 L^{2s} - \cdots - \alpha_P L^{Ps} \tag{3}$$

$$\Theta_q(L) = 1 + \theta_1 L + \theta_2 L^2 + \cdots + \theta_q L^q \tag{4}$$

$$B_Q(L^s) = 1 + \beta_1 L^s + \beta_2 L^{2s} + \cdots + \beta_Q L^{Qs} \tag{5}$$

SARIMA模型的构建流程如下:首先采用ADF检验(Augmented Dickey-Fuller Test)识别时间序列的平稳性,并采用趋势差和季节差将非平稳数据转换成平稳数据.然后根据上述平稳时间序列的自相关函数(ACF)和偏自相关函数(PACF),结合赤池信息准则(Akaike information criterion,AIC)选择p、q、P和Q的最优参数组及模型结构.

2.2.2　Prophet模型　Prophet模型的数学表达式见式(6).

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_i \tag{6}$$

式中,t为时间,$g(t)$为趋势项,$s(t)$为季节项(周期项),$h(t)$为表示节假日效应,ε_i为误差项.

趋势项$g(t)$模拟了时间序列在非周期上的变化趋势,包含非线性饱和增长即逻辑斯蒂曲线(式(7))和线性增长(式(8))两种函数.

$$g(t) = \frac{C(t)}{1 + \exp\left(-k + a(t)^T\delta\right)(t - (m + a(t)^T\gamma))} \tag{7}$$

$$g(t) = \left(k + a(t)^T\delta\right)t + (m + a(t)^T\gamma) \tag{8}$$

式中,k为增长率,δ为速率调整,m为偏移参数,γ为连续变化点调整,$C(t)$为时间序列在任意时间点的预期容量.

季节项(周期项)$s(t)$考虑了时间序列数据中各种周期类型的变化趋势,采用傅里叶级数来表达周期规律,公式见式(9).

$$s(t) = \sum_{n=1}^{N}\left(a_n \cos\left(\frac{2\pi nt}{P}\right) + b_n \cos\left(\frac{2\pi nt}{P}\right)\right) \tag{9}$$

式中,a和b分别为观察到的季节性,P为一个周期的时间序列长度,N为周期个数.

节假日效应$h(t)$用于表述时间序列在某些特殊时间的变化规律,其表达式见式(10).

$$h(t) = \sum_{i=1}^{L}\kappa_i \cdot 1_{\{t \in D_i\}} \tag{10}$$

式中,D为过去和未来的节假日数据集,每个假期都包含一个参数κ_i.

Prophet模型的构建流程如下:根据数据的趋势性、周期性及节假日效应特征不断调整参数直至达到要求的预测精度,最后进行预测结果和评价指标的输出.表1列出了模型的主要参数及设定值.

表1　Prophet模型主要参数及设定值

Table 1　Main parameters and set values of the Prophet model

参数名称	设定值
对转折点拟合的灵敏度	0.05
转折点个数	25
年拟合度	FALSE
周拟合度	TRUE
日拟合度	FALSE
增长模式	“Linear”

2.2.3　LSTM模型　基于梯度的长短期记忆神经网络(LSTM)作为递归神经网络(RNN)的变体,由Sepp等首次提出(Hochreiter *et al.*,1997).它有效地避免了RNN中梯度消失和梯度爆炸的现象,并增强了时间序列的预测结果.模型的表达式分别见式(11)~(16).

$$i_t = \sigma(\boldsymbol{W}_{xi}x_t + \boldsymbol{W}_{hi}h_{t-1} + \boldsymbol{W}_{ci}c_{t-1} + b_i) \tag{11}$$

$$f_t = \sigma(\boldsymbol{W}_{xf}x_t + \boldsymbol{W}_{hf}h_{t-1} + \boldsymbol{W}_{cf}c_{t-1} + b_f) \tag{12}$$

$$c_t = f_t c_{t-1} + i_t \tanh(\boldsymbol{W}_{xc}x_t + \boldsymbol{W}_{hc}h_{t-1} + \boldsymbol{W}_{ci}c_{t-1} + b_c) \tag{13}$$

$$o_t = \sigma(\boldsymbol{W}_{xo}x_t + \boldsymbol{W}_{ho}h_{t-1} + \boldsymbol{W}_{co}c_{t-1} + b_0) \tag{14}$$

$$h_t = o_t \tanh(c_t) \tag{15}$$

$$y(t) = \boldsymbol{W}_{yh}\left(f\left(\boldsymbol{W}_{hx}y_{t-1} + \boldsymbol{W}_{hh}h_{t-1} + b_h\right) + b_y\right) \tag{16}$$

式中，i_t为输入门；f_t为遗忘门；c_t为t时刻的记忆细胞状态；o_t为输出门；h_t为t时刻的隐藏状态；$\boldsymbol{W}_{xc}$、$\boldsymbol{W}_{xi}$、$\boldsymbol{W}_{xf}$和$\boldsymbol{W}_{xo}$分别为连接输入信号的权重矩阵；$\boldsymbol{x}_t$、$\boldsymbol{W}_{hc}$、$\boldsymbol{W}_{hi}$、$\boldsymbol{W}_{hf}$和$\boldsymbol{W}_{ho}$分别为连接隐藏层输出信号的权重矩阵；$\boldsymbol{h}_t$、$\boldsymbol{W}_{ci}$、$\boldsymbol{W}_{cf}$和$\boldsymbol{W}_{co}$为连接神经元激活函数的对角矩阵；b_i、b_c、b_f、b_0、b_h和b_y代表偏差矢量；σ为激活函数（本研究为sigmoid函数）；$y(t)$为总叶绿素a时间序列预测值；$\boldsymbol{W}_{hx}$、$\boldsymbol{W}_{hh}$和$\boldsymbol{W}_{yh}$分别为输入-隐藏权重矩阵、隐藏-隐藏权重矩阵和隐藏-输出权重矩阵；$\boldsymbol{y}_{t-1} = \left(y_{t-1}, y_{t-2}, \cdots, y_{t-d}\right)'$为包含时间步长序列的向量.

LSTM模型的建模流程如下：在训练模型前将数据集划分为前70%的训练数据集和后30%的测试数据集，并从训练集中抽取20%的数据作为验证集，用于对模型的结构和参数进行调整.为提高模型的收敛速度和预测精度，通过min-max标准化方法将原始数据进行归一化处理并转化为[0,1]之间的数值.模型的主体结构为Stacked-LSTM，两层LSTM的神经元个数分别为100和50，并在各层间增加随机失活（Dropout）层以防止模型的过拟合.激活函数和优化算法分别选择Sigmoid及RMSProp.后者是对梯度下降方法的改进，可改善深度学习方法中学习率明显下降且过早结束的问题，较适合处理非平稳的藻类数据.模型的损失函数评价指标为均方误差（Mean Squared Error，MSE），每次迭代的批大小（Batch size）为50，训练步数（Epoch）为200.使用TensorFlow框架下的早停法（Early Stopping）对验证集误差进行监测，有效避免模型的过拟合.基于LSTM模型的总叶绿素a预测流程见图2.

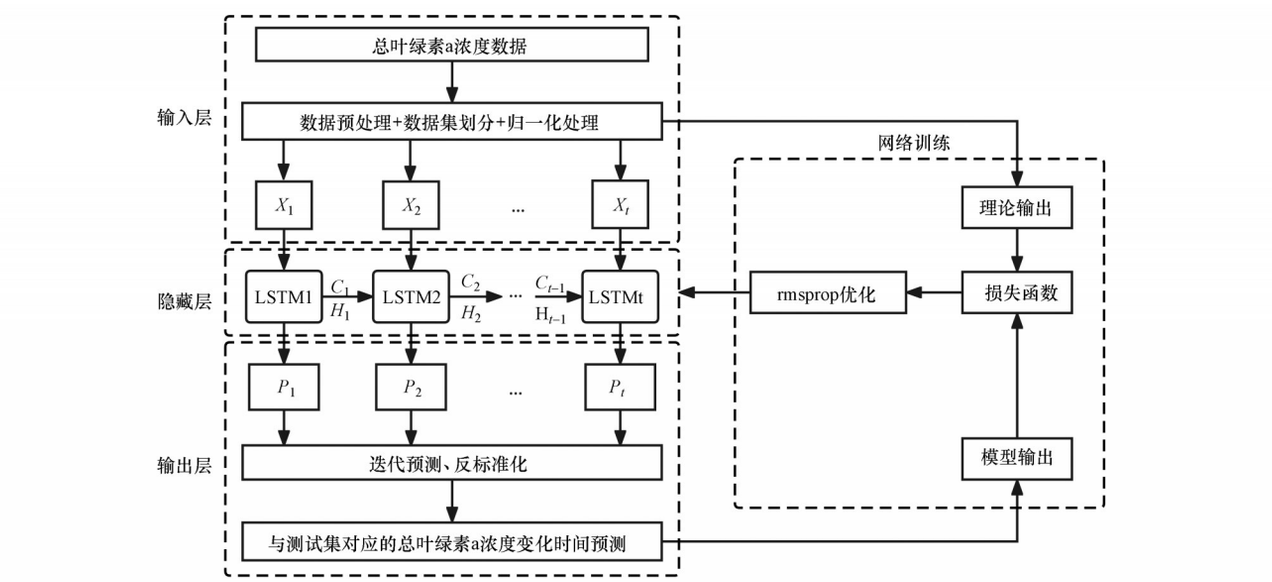

图2 基于LSTM模型的总叶绿素a预测流程

Fig.2 Prediction routine of total chlorophyll a based on the LSTM model

2.3 藻华表征

通常采用叶绿素a浓度阈值判别藻类水华的形成，但目前尚无统一标准（黄群芳等，2022）.参考相关研究（张艳会等，2016），并对江东库区原始数据进行分析，发现总叶绿素a的日平均浓度高于15.0 μg·L^{-1}时，水体中溶解氧饱和度与浓度的关系出现拐点且两者显著相关（图3）.因此，本文定义日总叶绿素a平均浓度高于15.0 μg·L^{-1}时为藻华风险日（简称藻华日），若日平均浓度低于或等于该数值，则定义为非藻华日.一个或多个连续藻华日形成的时间序列称为一个藻华风险周期.在该定义下，2017年7月—2020年6月，藻华风险周期的平均藻华日为2.54 d，其中持续7 d及以下的藻华风险周期占96%.因此，采用未来7 d作为模型预测时长，基本反映本研究库区的藻华预测预警的实际需要.

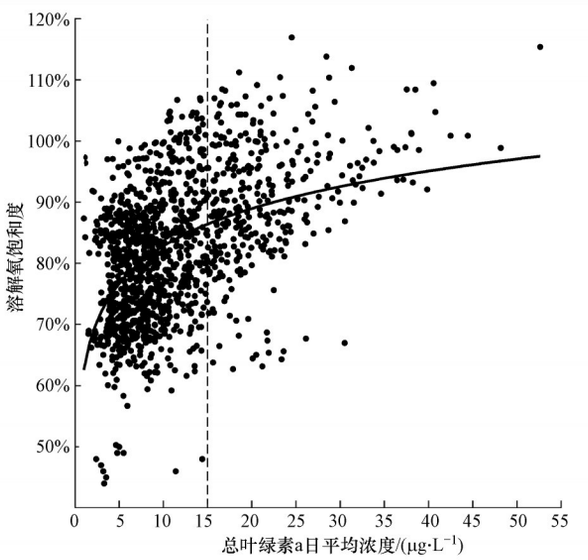

图3　溶解氧饱和度随总叶绿素a日平均浓度的变化情况(2017年7月—2020年6月)

Fig.3　Dissolved oxygen saturation against daily average concentration of total chlorophyll-a in July 2017 to June 2020

2.4　模型性能评估与统计检验

采用RMSE(Root Mean Squared Error)作为判断模型预测效果优劣的依据,其表达式见式(17).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(y_t - y_p\right)^2} \tag{17}$$

式中,n为样本总数,y_t为总叶绿素a浓度实际值,y_p为总叶绿素a浓度预测值.

为进一步比较不同实验组(3种模型、不同的输入时间长度、不同的输入数据频率)的预测结果,使用Wilcoxon符号平均秩检验对组间差异进行比较.检验步骤如下:①将第2组样本的各个观察值减去第1组样本对应的观察值,若差值为正记为正号,若差值为负记为负号,若差值等于0则删除此个案,样本数n也相应减少;②保留差值数据,根据其绝对值大小按升序排序,并求出相应的秩;③分别计算符号为正号的秩和W_+、符号为负号的秩和W_-,以及正号平均秩、负号平均秩;④根据式(18)计算Z值;⑤根据检验统计量计算相伴概率值,并与设定的显著性水平进行比较,最终作出检验判断.

$$Z = \frac{W - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{18}$$

式中,$W = \min\left(W_+,\ W_-\right)$,$n$为删除差值等于0的个案以后的样本量.

3　结果(Results)

3.1　藻华特征

2017年7月1日—2020年6月30日3年九龙江江东库区总叶绿素a浓度随时间变化过程如图4所示,其小时平均和日平均浓度的波动幅度较大.根据定义可知,3年的藻华日为254 d,占总日数的23.2%.其中,2017年7月—2018年6月共有藻华日91 d,2018年7月—2019年6月共有藻华日77 d,2019年7月—2020年6月共有藻华日86 d.

3.2　3种时间序列模型预测效果评估

使用3种时间序列模型分别进行总叶绿素a日均浓度预测,采用输入前7 d均值,输出后7 d均值的模式.将LSTM模型测试集的预测结果(RMSE)与同时段SARIMA和Prophet模型的预测结果进行逐日比较(图5a),3种模型预测值的RMSE均随时间呈现增加趋势,其中,Prophet模型的RMSE最大,LSTM模型的RMSE最小.3个模型7 d预测值的平均RMSE有显著差异(图5b),LSTM模型预测的RMSE显著低于SARIMA模型和Prophet模型,3个模型的平均RMSE分别为6.4、7.3和12.8 μg·L^{-1}.

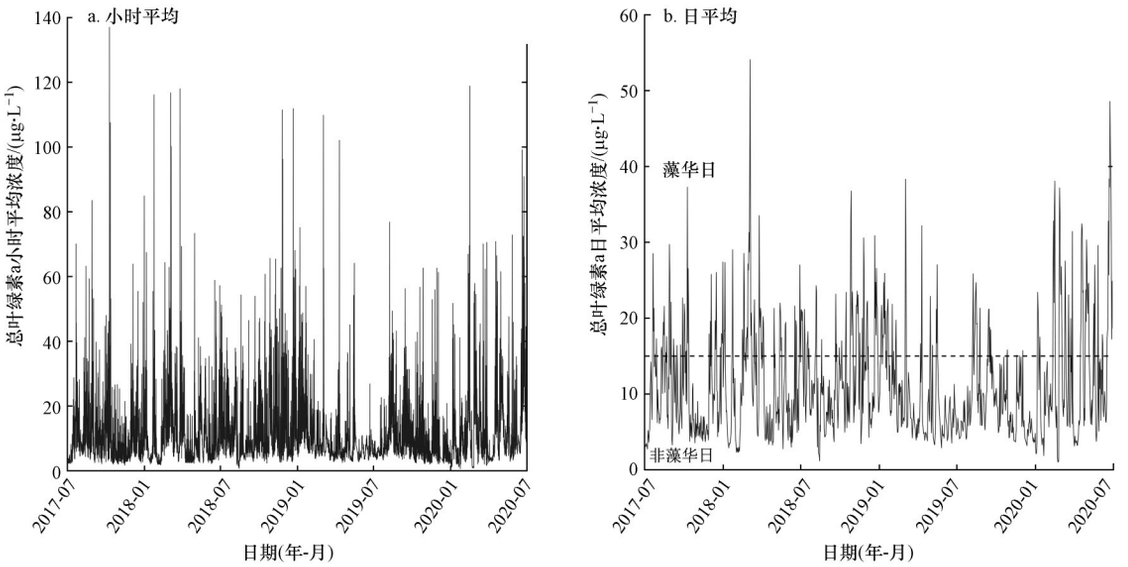

图4 九龙江江东库区总叶绿素a浓度随时间变化过程(2017年7月1日—2020年6月30日)

Fig.4 Temporal variation of total Chl-a concentration in the Jiangdong reservoir of Jiulong River (July 1st, 2017 to June 30th, 2020)

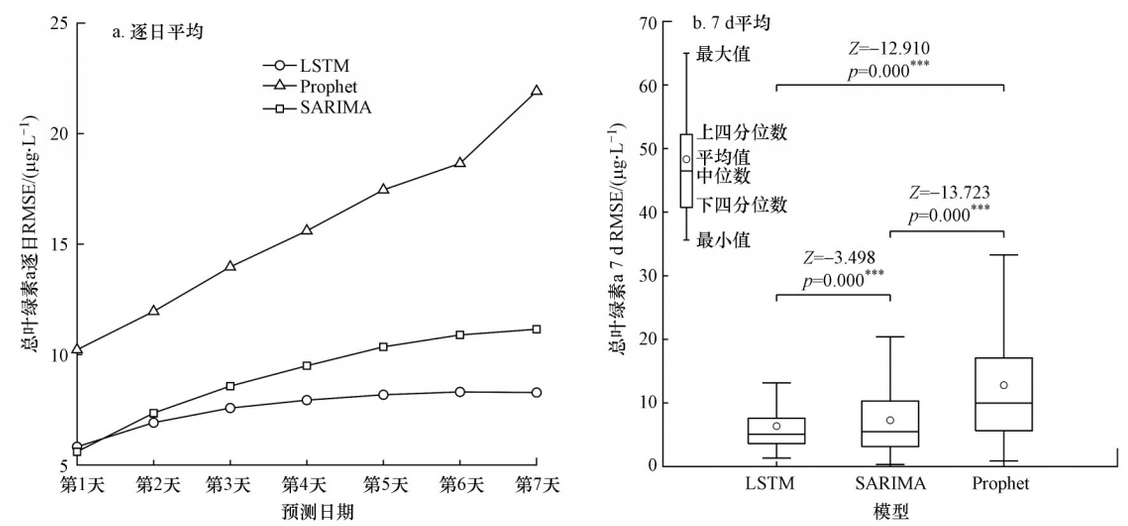

图5 3种模型预测未来7 d总叶绿素a的RMSE比较

Fig.5 Comparison of RMSE of predicted total Chl-a by three models in the next seven days

3.3 输入数据时间长度对LSTM模型预测效果的影响

调整LSTM模型的输入数据时间长度(7、15、30、150 d),观察总叶绿素a预测结果的差异.当输入数据时间长度为7 d时,模型的逐日RMSE达到最低;而随着输入数据时间长度的增加,逐日RMSE均呈增加趋势(图6a).模型7 d预测的平均RMSE也呈上升趋势(图6b).如果只统计藻华日,也是输入数据时间为7 d时的预测RMSE最低(图7).总体而言,当模型输入数据时间长度为7 d时,预测效果最佳.

3.4 输入数据时间频率对LSTM模型预测效果的影响

在固定输入数据时间长度为7 d时,分别采用小时频率和日频率两种数据进行300次的预测实验.采用小时频率的数据进行预测时,将预测值和真实值进行日平均后再计算RMSE进行比较.结果显示,如果只统计非藻华日,小时频率的7 d预测RMSE显著低于日频率(p=0);如果只统计藻华日,两者的7 d预测RMSE并不存在显著差异(p=0.107).

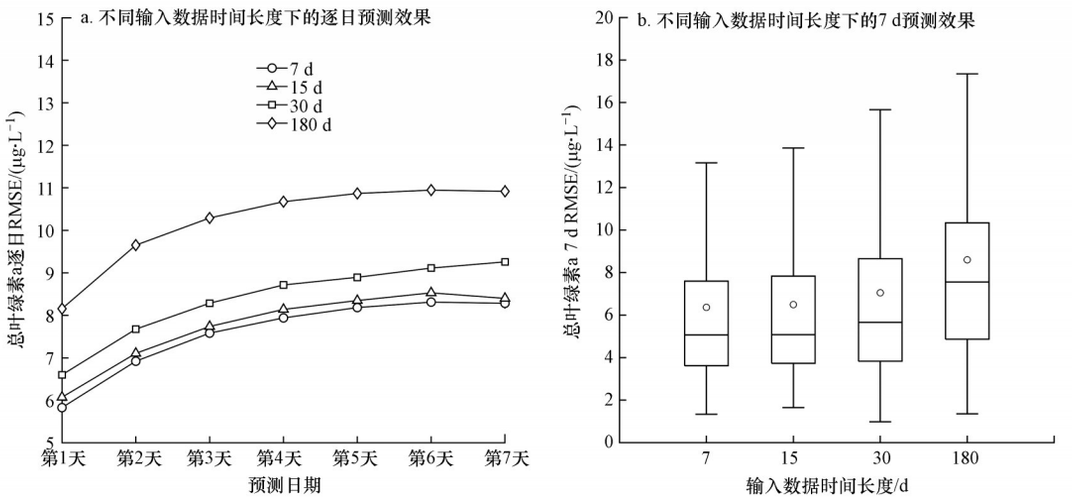

图6　输入数据时间长度对LSTM模型预测总叶绿素a的RMSE的影响

Fig.6　Impact of input data length on prediction RMSE of total Chl-a by the LSTM model

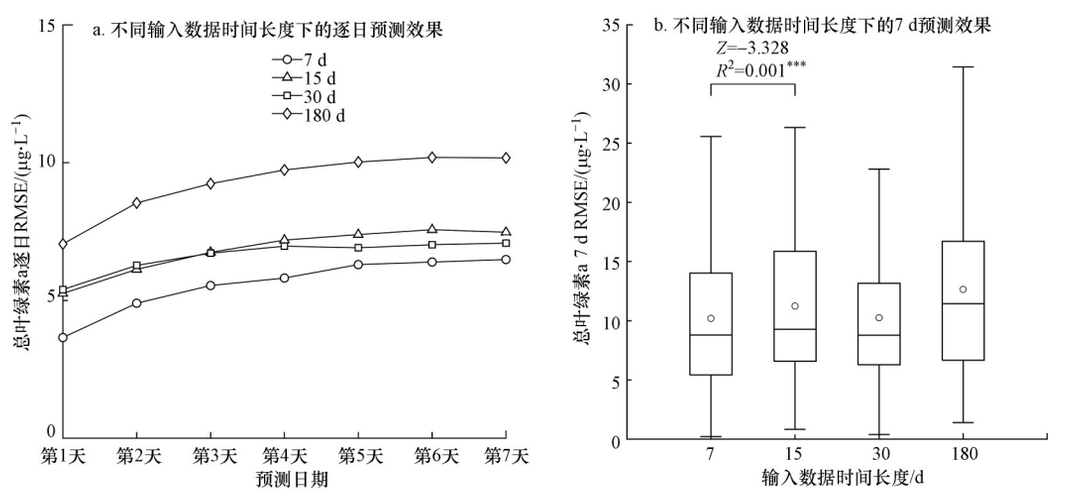

图7　输入数据时间长度对LSTM模型预测藻华日总叶绿素a的RMSE的影响(藻华日预测)

Fig.7　Impact of input data length on prediction RMSE of total Chl-a by the LSTM model (prediction of algal bloom day)

为更明晰地展示模型的逐日预测效果,绘制了两种数据频率下LSTM模型的预测值与实测值对比图(图8).基于日频率数据的藻华日的逐日预测RMSE均低于基于小时频率的预测RMSE,但非藻华日的RMSE除了第1 d外并无显著差异.总体而言,日频率的预测效果优于小时频率的预测.

4　讨论(Discussion)

4.1　时间序列模型相对传统藻类监测预警方法的优势

传统藻类监测预警主要基于过程(Process-based)或机理(Mechanism-driven)模型,比如可以模拟浮游植物生长和消亡的生态动力学模型,但需要大量的环境和藻类数据,模型参数率定困难(陈声威,2014),大部分河流库区并不具备这样的数据条件,很难进行部署应用(Huang *et al.*,2020).比如,环境流体动力学模型(EFDC)常用于库区藻华预测,但因为藻华爆发机理复杂,影响因子多,导致藻类高值预测效果较差(Kim *et al.*,2017).相比之下,在线监测技术与时间序列模型相结合可有效弥补传统藻类监测预警方法的局限性.本研究采用的时间序列模型相对于传统藻类监测预警方法具有明显优势.作为一种数据驱动(Data-driven)模型,时间序列模型具有灵活性较强的数学结构,通过机器学习有效识别历史数据的变化规律(Parmezan *et al.*,2019).时间序列模型能更为清晰地反映水质特征和未来变化趋势(陈能汪等, 2021).

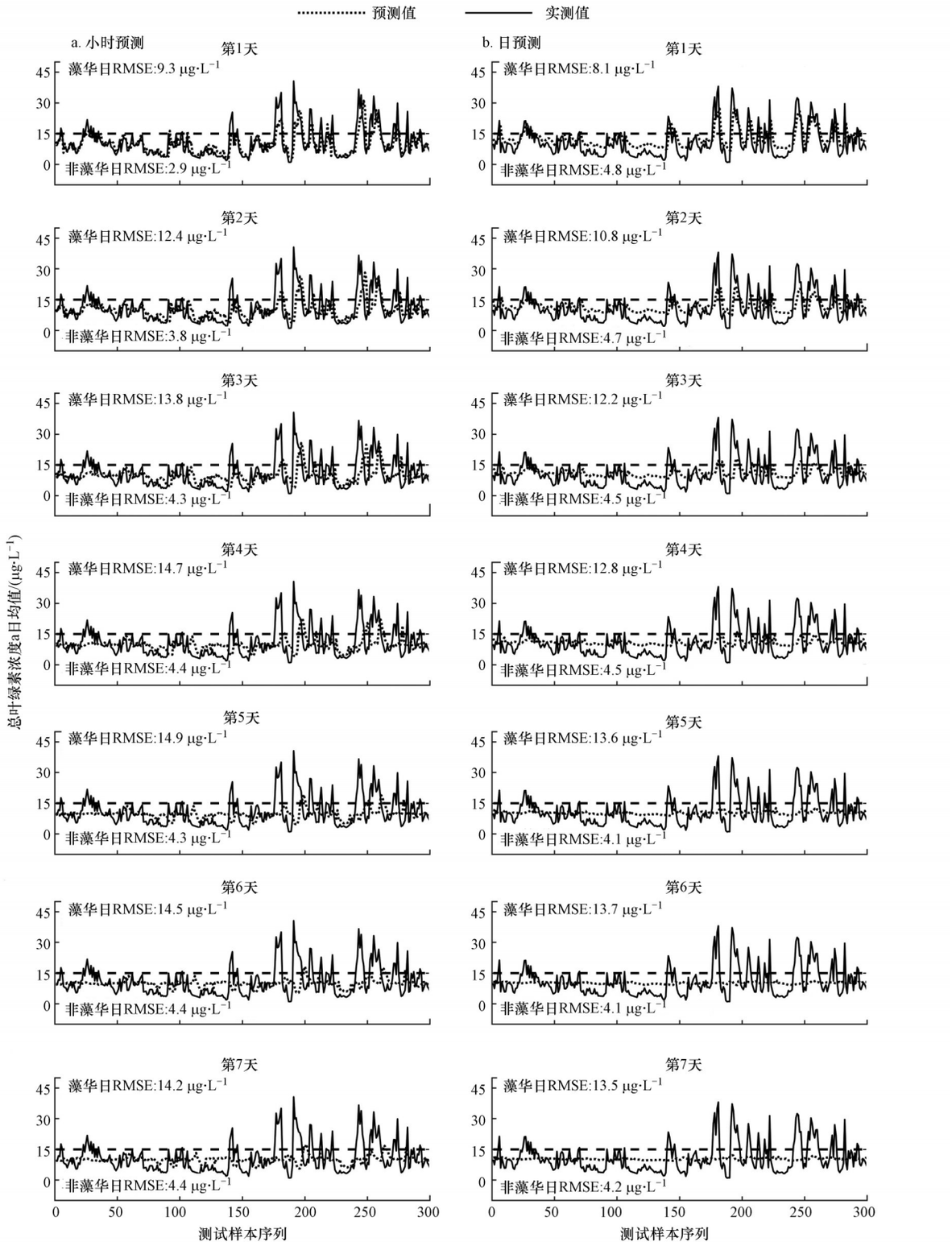

图8 基于小时频率和日频率数据的LSTM模型预测未来7 d总叶绿素a的结果比较

Fig.8 Prediction of total Chl-a in future seven days by LSTM based on hourly and daily data

4.2 LSTM模型相对SARIMA模型和Prophet模型的优势

研究发现,LSTM模型在预测总叶绿素a时间序列上相对于其他两种模型具有显著优势(图5).作为一种深度学习技术,LSTM模型的梯度下降原理及“迭代”优化算法有助于构建最适合数据分布特征的模型,能

较好地识别数据中存在的非线性趋势.由于藻类水华受物理、化学及生物因素的共同影响,机理十分复杂(吴娟等,2020),时间序列数据的随机性较强,这导致了SARIMA的预测效果欠佳(Wang *et al.*,2021).而Prophet模型不考虑残差间的自相关,在模型训练时容易造成欠拟合,无法准确学习藻类数据复杂的变化特征,从而影响预测效果(Guo *et al.*,2021).SARIMA和Prophet作为两种自回归模型,对非线性数据的识别和预测能力相比简单统计模型已有进步,但仍有较大的提升空间.

在面对水质指标的非线性和非平稳性特征时,常规的统计学模型往往难以进行准确预测(陈能汪等,2021),而基于深度学习框架的智能算法—LSTM模型作为一种循环神经网络(Recurrent Neural Network,RNN),会收集隐藏层中不断延长的序列数据,基于默认参数组能更有效地捕捉长时间序列的相关性(Parmezan *et al.*,2019; Yussof *et al.*,2021).因此,该模型可高效学习藻类时间序列历史数据动态变化的内在规律,并确定过去趋势对未来预测产生的可能影响(Saboe *et al.*, 2021),在应用于机理不明确的高维非线性系统时具有明显优势,在藻类预测相关研究中表现出色(Tian *et al.*, 2019;Rostam *et al.*,2021;Yussof *et al.*,2021),为藻华预测和预警提供了有效解决方案.

4.3　输入数据时间长度和频率对单变量LSTM模型预测效果的影响

本研究发现LSTM模型的总叶绿素a预测RMSE随输入数据时间长度的增加而逐渐上升(图6~7).根据机器学习理论,较长的历史时间序列会提升模型的复杂度,容易造成过拟合使得预测效果变差(Zhu *et al.*,2019).过于长期的时间依赖关系会增加LSTM的预测误差(Yang *et al.*,2019),这也印证了本研究的结果.

本研究发现两种频率数据的藻华日预测结果差异不显著,而非藻华日的预测结果则差异显著.对于藻华日,LSTM模型的预测误差不依赖于数据的频谱特征(汪凯翔等,2020);对于非藻华日,由于藻类浓度在多数情况下处于较低水平,小时频率实验可充分学习非藻华的特征.藻华日和非藻华日预测的第1~7 d,日频率实验的逐日RMSE均低于小时频率实验(图8).总体而言,日频率数据的藻华预测效果优于小时频率数据,这得益于较低的频率能使LSTM模型更高效地学习藻类水华形成及消亡的生态学特征.有研究表明,增加训练数据的频率对预测效果的改善有限(孙红等, 2021;蔡明昕等, 2021),且数据量大需要较长的训练时间,对硬件设备的要求较高,技术的适用性较差.因此,采用日频率数据的预测模式,可为藻华早期预警提供有效的技术支持.

综上所述,LSTM模型是一种比较理想的藻华预测技术.相比基于过程机理的生态动力学模型,LSTM模型所需的参数少,计算速度快,且不易出现因异常数据所引起的局部最优问题(于家斌等,2018).作为一种循环神经网络,LSTM模型能够较好地解决时间序列数据过长时容易出现的梯度消失或梯度爆炸问题,并较好地识别和学习藻类数据的非线性变化特征.当然,在实际应用LSTM模型进行藻华预测时仍需考虑是否具备稳定可靠的时间序列监测数据,数据质量决定了模型预测结果的优劣.若时间序列数据较短,模型有较高的过拟合(Overfitting)风险,可能造成其在测试集的表现较差(Parmezan *et al.*, 2019).不同的随机权重初始化也会对模型的表现产生影响(Yussof *et al.*, 2021).此外,对于大样本数据的训练,该模型仍有一定的改进空间,可加入卷积层,逐步从输入层移动到各个输出层进行特征提取,在显著减少训练时间的同时降低过拟合的可能性(陈能汪等,2021).未来可进一步优化模型架构,采用更高效的参数优化机制,如融合注意力机制,以提升模型对藻类水华的预测效果.

5　结论(Conclusions)

1) 在线监测技术与时间序列模型相结合可有效弥补传统藻类监测预警方法的局限性,时间序列模型可组成灵活性较强的数学结构,有效识别历史数据变化规律等重要信息,能更好地反映水质特征和未来变化趋势.

2)基于深度学习框架的LSTM模型,因其独特的迭代优化算法及对藻类非线性数据规律有较强的适应性,适用于机理复杂的非线性生态环境系统,其对未来7 d总叶绿素a浓度的预测效果显著优于SARIMA和Prophet这两种自回归模型.

3)改变LSTM模型输入数据长度和频率,发现输入数据长度(7~180 d)会在一定程度上影响模型预测效

果,较长的历史时间序列易造成模型的过拟合,使得预测效果降低,输入数据长度为7 d时,未来7 d的藻华预测性能最佳;输入数据频率对预测效果存在一定的影响,对于藻华日预测,小时频率与日频率数据对预测结果无显著影响,而在逐日预测实验中,基于日频率数据的预测效果更优.对于非藻华日,小时频率数据能为LSTM模型提供更多的特征学习,小时频率的预测效果优于日频率数据.总体而言,本文建立的基于单变量时间连续数据的LSTM模型具备河流库区藻华短期预测能力,可为九龙江流域水生态管理、库区有害藻华防控提供有力的技术支持,以保障供水安全,助力流域水生态治理能力现代化.

致谢:感谢福建省九龙江北溪水资源调配中心对本研究的支持.

参考文献(References):

Aditya S, Darmawan W, Nadia B U, *et al*. 2021. Time series analysis and forecasting of coronavirus disease in Indonesia using ARIMA model and PROPHET[J]. Procedia Computer Science, 179: 524-532

Atzori F, Jerono P, Schaum A, *et al*. 2021. Identification of a cell population model for algae growth processes[J]. IFAC-PapersOnLine, 54(7): 132-137

Bouktif S, Fiaz A, Ouni A, *et al*. 2020. Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting[J]. Energies, 13(2): 391-413

蔡明昕,孙晶,王斌.2021.多角度语义轨迹相似度计算模型[J],计算机科学与探索,15(9):1632-1640

Chen N, Hong H, Gao X. 2021. Securing drinking water resources for a coastal city under global change: Scientific and institutional perspectives[J]. Ocean & Coastal Management, 207: 104427-104434

陈能汪,余镒琦,陈纪新,等.2021.人工神经网络模型在水质预警中的应用研究进展[J].环境科学学报,41(12):4771-4782

陈能汪,章颖瑶,李延风.2010.我国淡水藻华长期变动特征综合分析[J].生态环境学报,19(8):1994-1998

陈声威.2014.水体富营养化预警模型研究现状和发展趋势[J].水利科技与经济,20(4):5-8+10

Daghighi A. 2017. Harmful algae bloom prediction model for Western Lake Erie using stepwise multiple regression and genetic programming[D]. USA: Cleveland State University.

Griffith A W, Gobler C J. 2020. Harmful algal blooms: A climate change co-stressor in marine and freshwater ecosystems[J]. Harmful Algae, 91: 101590-101601

Guo L, Fang W, Zhao Q, *et al*. 2021. The hybrid PROPHET-SVR approach for forecasting product time series demand with seasonality[J]. Computers & Industrial Engineering, 161: 107598-107611

Hochreiter S, Schmidhuber J. 1997. Long short-term memory[J]. Neural Computation, 9(8): 1735-1780

Huang C, Huang H, Li Y. 2019. A bidirectional LSTM prognostics method under multiple operational conditions[J]. IEEE Transactions on Industrial Electronics, 66(11): 8792-8802

Huang J, Zhang Y, Arhonditsis G B, *et al*. 2020. The magnitude and drivers of harmful algal blooms in China's lakes and reservoirs: A national-scale characterization[J]. Water Research, 181: 115902-116916

黄群芳,国超旋,李娜,等.2022.富春江库区高温热浪变化特征及对藻类水华潜在影响研究[J].环境科学研究, 35(2):530-539

Kim J, Lee T, Seo D. 2017. Algal bloom prediction of the lower Han River, Korea using the EFDC hydrodynamic and water quality model[J]. Ecological Modelling, 366: 27-36

Lee S, Lee D. 2018. Improved prediction of harmful algal blooms in four major South Korea's rivers using deep learning models[J]. International Journal of Environmental Research and Public Health, 15(7): 1322-1336

Li Y, Cao W, Su C, *et al*. 2011. Nutrient sources and composition of recent algal blooms and eutrophication in the northern Jiulong River, Southeast China [J]. Marine Pollution Bulletin, 63(5/12): 249-254

Ly Q V, Nguyen X C, Lê N C, *et al*. 2021. Application of machine learning for eutrophication analysis and algal bloom prediction in an urban river: A 10-year study of the Han River, South Korea[J]. Science of the Total Environment, 797: 149040-149053

Ma J, Deng J, Qin B, *et al*. 2013. Progress and prospects on cyanobacteria bloom-forming mechanism in lakes[J]. Acta Ecologica Sinica, 33(10): 3020-3030

Park Y, Cho K H, Park J, *et al*. 2015. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuarine reservoirs, Korea[J]. Science of The Total Environment, 502: 31-41

Parmezan A R, Souza V M, Batista G E. 2019. Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model[J]. Information Sciences, 484: 302-337

Rabbani M B, Musarat M A, Alaloul W S, *et al*. 2021. A comparison between seasonal autoregressive integrated moving average (SARIMA) and exponential smoothing (ES) based on time series model for forecasting road Accidents[J]. Arabian Journal for Science and Engineering, 46(11): 11113-11138

Rahman A T, Hosono T, Kisi O, *et al*.2020.A minimalistic approach for evapotranspiration estimation using the Prophet model[J].Hydrological Sciences Journal,65(12):1994-2006

Rostam N A, Malim N H, Abdullah R, *et al*.2021.A complete proposed framework for coastal water quality monitoring system with algae predictive model [J].IEEE Access,9:108249-108265

Saboe D, Ghasemi H, Gao M M, *et al*.2021.Real-time monitoring and prediction of water quality parameters and algae concentrations using microbial potentiometric sensor signals and machine learning tools[J].Science of the Total Environment,764:142876-142883

Sangiorgio M, Dercole F.2020.Robustness of LSTM neural networks for multi-step forecasting of chaotic time series[J].Chaos, Solitons & Fractals,139: 110045-110056

Shin Y, Kim T, Hong S, *et al*.2020.Prediction of chlorophyll-a concentrations in the Nakdong River using machine learning methods[J].Water,12(6): 1822-1839

孙红,黎铨祺,赵娜.2021.基于双层树状支持向量机的观点挖掘与倾向分析[J].智能计算机与应用,11(3):44-47

Taylor S, Letham B.2017.Forecasting at scale[R].PeerJ Preprints

Tian W, Liao Z, Wang X.2019.Transfer learning for neural network model in chlorophyll-a dynamics prediction[J].Environmental Science and Pollution Research,26(29):29857-29871

Toharudin T, Pontoh R S, Caraka R E, *et al*.2021.Employing long short-term memory and Facebook prophet model in air temperature forecasting[J]. Communications in Statistics - Simulation and Computation:1-24,DOI:10.1080/03610918.2020.1854302

汪凯翔,黄清华,吴思弘.2020.长短时记忆神经网络在地电场数据处理中的应用[J].地球物理学报,63(8):3015-3024

Wang L, Wang X, Jin X, *et al*.2017.Analysis of algae growth mechanism and water bloom prediction under the effect of multi-affecting factor[J].Saudi Journal of Biological Sciences,24(3):556-562

Wang X, Tian W and Liao Z.2021.Statistical comparison between SARIMA and ANN's performance for surface water quality time series prediction[J]. Environmental Science and Pollution Research,28(25):33531-33544

Wang Y, Xu C, Zhang S, *et al*.2019.Development and evaluation of a deep learning approach for modeling seasonality and trends in hand-foot-mouth disease incidence in mainland China[J].Scientific Reports,9(1):8046-8060

吴娟,朱跃龙,金松,等.2020.三种机器学习模型在太湖藻华面积预测中的应用[J].河海大学学报(自然科学版),48(6):542-551

Yang B, Sun S, Li J, *et al*.2019.Traffic flow prediction using LSTM with feature enhancement[J].Neurocomputing,332:320-327

Yang X, Xu H, Shu H, *et al*.2021.Service component recommendation based on LSTM[J].International Journal of Embedded Systems,14(2):201-209

杨昆,罗毅,徐玉妃,等.2016.基于无线传感器网络与GIS的蓝藻水华爆发动态监测与模拟[J].农业工程学报,32(24):197-205

于家斌,尚方方,王小艺,等.2018.基于遗传算法改进的一阶滞后滤波和长短期记忆网络的蓝藻水华预测方法[J].计算机应用,38(7):2119-2123+ 2135

Yussof F N, Maan N, Md Reba M N.2021.LSTM Networks to improve the prediction of harmful algal blooms in the west coast of Sabah[J].International Journal of Environmental Research and Public Health,18(14):7650-7663

于洋,彭福利,孙聪,等.2017.典型湖泊水华特征及相关影响因素分析[J].中国环境监测,33(2):88-94

张艳会,李伟峰,陈求稳.2016.太湖水华程度及其生态环境因子的时空分布特征[J].生态学报,36(14):4337-4345

Zhu Y, Zhang W, Chen Y, *et al*. 2019. A novel approach to workload prediction using attention-based LSTM encoder-decoder network in cloud environment[J].EURASIP Journal on Wireless Communications and Networking,2019(1):274-291

Zohdi E, Abbaspour M.2019.Harmful algal blooms (red tide): A review of causes, impacts and approaches to monitoring and prediction[J].International Journal of Environmental Science and Technology,16(3):1789-1806