

吴畅, 崔婧嫒, 黄帆, 等. 基于 STL-MIKE-LSTM 联合建模的陈行水库盐度快速预测[J]. 净水技术, 2025, 44(5): 57–68, 186.

WU C, CUI J Y, HUANG F, et al. Rapid prediction of salinity for Chenhang Reservoir based on joint modeling of STL-MIKE-LSTM [J]. Water Purification Technology, 2025, 44(5): 57–68, 186.

基于 STL-MIKE-LSTM 联合建模的陈行水库盐度快速预测

吴 畅¹, 崔婧嫒², 黄 帆³, 宋辰煜¹, 张 晟¹, 赵蓬勃¹, 张海平^{1,*}

(1. 同济大学环境科学与工程学院, 上海 200092; 2. 浙江水利水电学院水科学研究院, 浙江杭州 310018; 3. 宜宾学院质量管理与检验检测学部, 四川宜宾 644000)

摘 要 【目的】 该研究欲探究使用神经网络模型替代湖泊水库非线性水文水质过程机理模型的可行性, 实现快速计算在咸潮入侵期间不同取水条件下的输水水质。【方法】 该研究提出了一种基于联合建模的氯化物输水浓度预测方法。该方法使用周期趋势分解算法(STL)分解提取取水浓度波动特征, 基于高斯噪声进行数据增强以构建合成样本, 并生成模拟工况方案, 交由 MIKE 21 机理模型计算, 最终将计算结果用于训练长短期记忆(LSTM)神经网络模型。【结果】 该研究将该方法施用于陈行水库, 结果如下: (1) 探究不同参数下 STL 分解的时序特征提取效果, 发现选择周期参数(n_p)=12 使 STL 分解的效果较好; (2) 对比不同隐藏层神经元个数与预测步长下 LSTM 模型的预测效果, 发现预测效果随神经元个数增加先升后降, 随预测步长增加持续下降, 选取神经元个数为 128 与预测步长为 24 h 的总体效果较好; (3) 对比不同结构神经网络的预测能力, 发现 LSTM 在预测集预测效果最佳[均方根误差(RMSE)=0.13 mg/L, 平均相对误差(MRE)=0.04, 纳什效率系数(NSE)=0.96]; (4) 使用实际实测数据与预报数据验证 LSTM 模型, 发现其对预测氯化物输水浓度具备较高精度(RMSE=0.29 mg/L, MRE=0.09, NSE=0.58), 且所需算力与时间远低于 MIKE 21 机理模型。【结论】 该研究提出的湖泊水库氯化物输水浓度预测方法经验证同时具备较高计算精度与速度, 能够替代机理模型为水库管理者应对咸潮入侵提供快速决策支持。

关键词 咸潮 长短期记忆(LSTM) 周期趋势分解算法(STL) 数据增强 MIKE 21

中图分类号: X524 **文献标志码:** A **文章编号:** 1009-0177(2025)05-0057-13

DOI: 10.15890/j.cnki.jsjs.2025.05.007

Rapid Prediction of Salinity for Chenhang Reservoir Based on Joint Modeling of STL-MIKE-LSTM

WU Chang¹, CUI Jingyuan², HUANG Fan³, SONG Chenyu¹, ZHANG Sheng¹, ZHAO Pengbo¹, ZHANG Haiping^{1,*}

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

2. Institute of Water Sciences, Zhejiang University of Water Resources and Electric Power, Hangzhou 310018, China;

3. Faculty of Quality Management and Inspection, Yibin University, Yibin 644000, China)

Abstract [Objective] The study aims to probe into the feasibility of substituting neural network models for the mechanism models based on nonlinear hydrological and water quality processes for lakes and reservoirs, so that the output water quality under different water intake conditions can be calculated rapidly during salt tide. [Methods] The study proposed a novel method for predicting chloride output water concentration for lakes and reservoirs. In this method, seasonal-trend decomposition using LOESS (STL) was used first to extract the characteristics of intake water concentration data. Then, data augmentation based on Gaussian noise was applied to synthesize simulated data. After that, simulated operation conditions were constructed and calculated by MIKE 21 mechanism model. Finally, the calculation result were used to train an long-short term memory(LSTM) neural network model. [Results] The method proposed was applied to Chenhang Reservoir and it turned out that: (1) By comparing the result of STL decomposition under

[收稿日期] 2024-12-14

[基金项目] 上海城投水务(集团)有限公司科研项目: 基于在线数据联动的陈行水源地水质预报及应对措施研究(KYYS220001); 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] 吴畅(2000—), 男, 硕士研究生, 研究方向为水环境模型, E-mail: 1850023@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。

different parameters, it turned out that choosing periodicity parameter(n_p)=12 could lead to a better performance; (2) By comparing the prediction performances of the LSTM model under different numbers of hidden layer neurons and prediction time steps, it was found that the prediction performance went up at first and then went down as the number of neurons increased, and the performance continued to go down as the prediction time step increase. The result showed that selecting 128 neurons and a prediction time step of 24 hours had the best overall performance; (3) By comparing the prediction performances of neural networks with different structures, it was found that the LSTM performed the best in the prediction set [root mean square error (RMSE) was 0.13 mg/L, mean relative error (MRE) was 0.04, Nash-sutcliffe efficiency coefficient (NSE) was 0.96]; (4) According to the evaluation of established LSTM model using online monitoring and forecast data, it was demonstrated that the LSTM model indeed had high accuracy in predicting the outlet concentration (RMSE was 0.29 mg/L, MRE was 0.09, NSE was 0.58), and the computility and time required were far lower than those of the MIKE 21 mechanism model. [**Conclusion**] The method to predict the delivery water concentration of chloride that is proposed in this study has been verified to have both high computational accuracy and speed, and can replace the mechanism model to provide rapid decision-making support for reservoir managers in response to salt tide.

Keywords salt tide long-short term memory (LSTM) seasonal-trend decomposition using LOESS (STL) data augmentation MIKE 21

每年 10 月至翌年 4 月的冬春枯水季节以及部分流域性极端干旱灾害时的汛期,长江入海口易发严重的咸潮入侵事件。以 2022 年为例,此年发生了较长持续时间的极端干旱灾害,使得长江流域出现了“汛期反枯”的现象,咸潮入侵较之往年显著提前且强度明显提高^[1]。咸潮入侵期间的长江口南支正面入侵或北支咸水倒灌会严重影响长江口沿岸包括陈行水库、青草沙水库等重要饮用水水源地的原水供应,对周围省市居民饮用水安全造成严重威胁^[2]。

在咸潮入侵期间,水库通常需要在确保供应原水的氯化物质量浓度不超过《生活饮用水卫生标准》(GB 5749—2022)中相应要求(≤ 250 mg/L)的同时最大限度取水,以寻求在尽可能缓解供水压力的同时,避免水库堤坝承受过大的内外压力差^[3]。因此,通常需要拟定多个不同的取水方案,依托于模型计算确定较优的方案用以实施。MIKE 21 是丹麦水力学研究所(DHI)研制的二维水动力-水质耦合机理模型,其由于功能强大、模拟精度高而被广泛运用于模拟河口、湖泊、海湾以及近岸区域的水流、波浪、泥沙及环境场,并已被众多研究运用于各类水质指标的模拟计算,并取得了较高的模拟精度^[4]。如夏妍等^[5]使用 MIKE 21 和 ECO Lab 构建了模拟精度优良的镜河流域二维水动力水质模型,其中对水位模拟的纳什效率系数(NSE)可达 0.78,相对误差(RE)达 5.75%。Parvaze 等^[6]使用 MIKE 11 NAM 模型对 Jhelum 河及其支流的非稳态流动进行了建模,建模结果显

示,在校验集与验证集上 NSE 分别可达 0.76 与 0.80,建模结果较好。

虽然现有的机理模型的发展较为成熟,但机理模型的预测精度通常依赖于较高的网格精度,而这使得水动力、水质方程的离散化迭代求解的计算量大,计算速度慢,也因此仅依赖机理模型难以很好地应对突发性事件^[7]。近年来,随着以神经网络为代表的深度学习技术的发展,数据驱动的建模方式愈发受到关注。由于神经网络的预测计算基于矩阵运算,因此计算速度远超机理模型。长短期记忆(long-short term memory, LSTM)网络是循环神经网络(recurrent neural network, RNN)中的一类,其由于具备同时学习数据间的短期依赖关系与长期依赖关系而被广泛运用于湖泊、水库等水体的各类水质指标预测。目前已有部分研究尝试使用机理模型计算结果训练 LSTM 神经网络模型,并取得了较好效果。如 Liang 等^[8]基于环境流体动力学模型(environmental fluid dynamics code, EFDC)模型的计算结果,使用 LSTM 在不同超参数与预测步长条件下建立了多个叶绿素 a 模型,结果显示,预测步长为 1 d 的模型平均 NSE 可达约 0.8。Zheng 等^[9]使用 EFDC 模型根据单测点数据对北运河流域的叶绿素 a 浓度建模,并使用 EFDC 计算结果训练 LSTM 模型,结果显示,在预测步长为 1~3 h 的条件下, NSE 可达 0.797~0.991,平均绝对百分误差(MAPE)可达 2.74%~13.16%,预测效果较好。神经网络模型强大的学习能力依赖于大量模型样本的训练。然而面对现实问题,往往难以收集到大量满意的数据样本,

从而保障模型的泛化能力。数据增强 (data augmentation, DA)^[10] 是 1 种常用于计算机视觉领域的数据集扩充技术,其能够使模型学习到更多的不变性,增加模型的泛化能力,同时减少过拟合的风险。数据增强技术的有效使用需要保证新增的数据集保留了原数据集样本的特征信息。对于时序数据,周期趋势分解算法 (seasonal-trend decomposition using LOESS, STL) 是一种以局部加权回归 (locally weighted regression, LOESS) 作为平滑方法的时间序列分解算法,其能够实现将时间序列数据分解为趋势项、季节项与残差项,因此具有良好的序列数据特征提取能力^[11]。本研究拟将 STL 分解运用于时序数据的数据增强,通过提取已有样本特征信息,构建合成数据样本以扩充数据集,用于神经网络模型的训练。

综上,此研究欲探究使用神经网络模型替代湖泊水库非线性水文水质过程机理模型的可行性,实现快速计算在咸潮入侵期间不同取水条件下的输水水质。为此,此研究提出了一种基于 STL-MIKE-LSTM 联合建模的水库氯化物输水浓度快速预测方法,该方法使用 STL 分解提取水库氯化物取水浓度波动特征,基于高斯噪声进行数据增强以构建合成数据样本,并生成模拟工况方案,将其交由 MIKE 21 机理模型计算,并将计算结果用于训练 LSTM 神经网络模型。此研究拟对比探究不同参数下 STL 分解的时序特征提取效果、LSTM 模型在不同预测步长与隐藏层神经元个数下的预测效果,以及不同隐藏层结构对神经网络模型预测能力的影响。并使用陈行水库的真实实测数据对该方法建立的 LSTM 模型进行验证。

1 材料与方法

1.1 研究区域

陈行水库位于上海市宝山区罗泾镇东部长江江堤外侧,地处长江口南支。水库于 1992 年 5 月建成投产,经增能后当前实际库容为 953 万 m^3 ,水域面积为 135 万 m^2 ,是上海地区重要的饮用水水源地之一。如图 1 所示,库区北部建有 2 座江心式取水泵站(取水口),泵站取水自北侧进入水库。库区东南角建有输水管道(输水口),向月浦、泰和等水厂供水。目前水库平均日供水量为 165 万 m^3 ,最大日供水量为 228 万 m^3 。

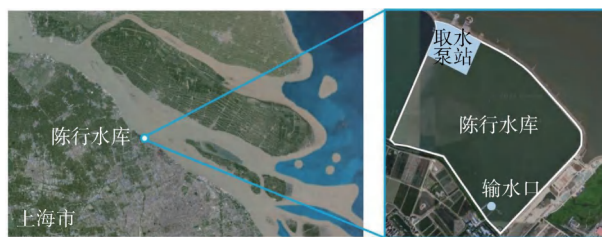


图 1 研究区域所在位置

Fig. 1 Location of the Study Area

1.2 模拟工况方案生成

1.2.1 模板数据选择

为克服数据量小带来的神经网络训练困难,本研究基于以下原则选取模板时间段,并将模板时间段对应的水库氯化物取水浓度在线监测数据作为模板数据,构建合成数据以拓展模型训练数据。

(1) 模板时间段内的在线监测数据完整无缺失。

(2) 模板时间段应涵盖 1 次完整的取水浓度涨落过程,且模板时间段前后取水浓度基本一致。

(3) 考虑到陈行水库实际遭遇超高浓度咸潮入侵时一般直接停止取水,本研究着重考虑峰值质量浓度小于 400 mg/L 的咸潮入侵。

根据以上原则,最终选取 2021 年—2023 年 3 次咸潮入侵时段为模板时间段,分别为 2021 年 3 月 1 日—2021 年 3 月 10 日、2022 年 12 月 24 日—2023 年 1 月 8 日以及 2023 年 1 月 24 日—2023 年 2 月 8 日,并依次编号为 I ~ III。模板数据的监测数据频率为 1 次/h。模板时间段与模板数据的选择如图 2 所示。

1.2.2 模板数据分解

由图 2 中选取的模板数据可知,在一次咸潮入侵过程中,取水浓度数据明显存在先升高再恢复的总体趋势,同时还存在高频周期性波动,其数据整体特征与 STL 分解原理较为契合。因此,在本研究中,为充分提取模板数据之中的趋势性与周期性特征,选择首先使用 STL 分解算法对模板数据进行处理。STL 分解算法使用内外两层循环迭代将原数据逐步分解为趋势项 (T)、季节项 (S) 与残差项 (R)。令 Y 表示模板数据(数据总长度记为 N), T^k 、 S^k 表示第 $k-1$ 轮内循环结束后得到的趋势项与季节项(其中 T^0 被初始化为 0), n_p 表示序列数据周期, n_s 、 n_1 、 n_i 分别表示不同的 LOESS 窗口宽度, n_i 、 n_o 分别

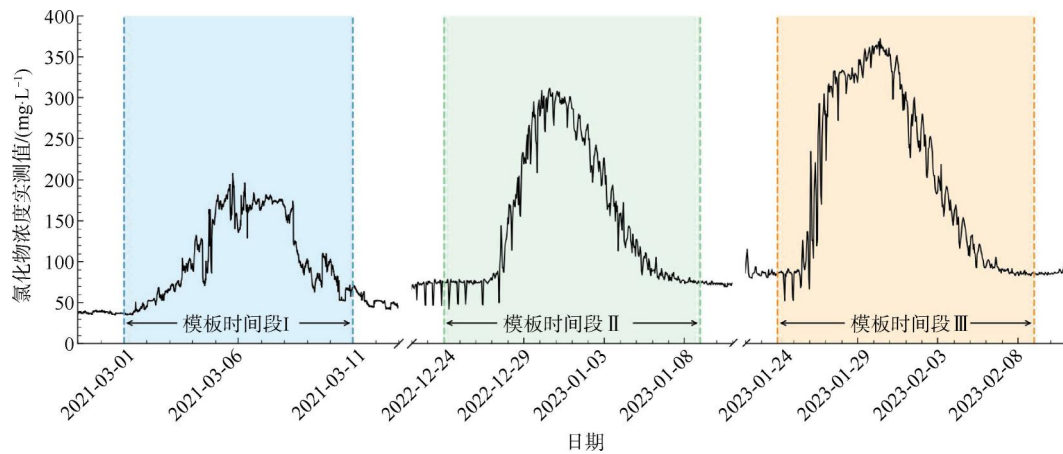


图2 模板时间段与数据

Fig. 2 Template Periods and Data

表示内循环与外循环次数,则 STL 分解算法的内循环的 6 个步骤如下^[11]。

(1) 去除趋势项:从模板数据去除上一轮内循环计算得到的趋势项,即计算 $Y-T^k$ 。

(2) 循环子序列平滑:将 $Y-T^k$ 以 n_p 为间隔选取数据作为 1 组,组成共 n_p 个循环子序列。对于每个循环子序列,作窗口宽度为 n_s 的 LOESS 平滑,并在两侧各衍生计算一个值,得到长度为 $N + 2n_p$ 的临时季节项,并记为 C^{k+1} 。

(3) 低通滤波:对 C^{k+1} 依次作 3 次窗口宽度分别为 $n_p, n_p, 3$ 的滑动平均,之后作窗口宽度为 n_l 的 LOESS 平滑,并将结果记为 L^{k+1} 。

(4) 去除循环子序列趋势项,获得季节项:计算 $S^{k+1} = C^{k+1} - L^{k+1}$ 。

(5) 去除季节项:从模板数据去除本轮内循环计算得到的季节项,即计算 $Y-S^{k+1}$ 。

(6) 趋势项平滑:对 $Y-S^{k+1}$ 作窗口宽度为 n_l 的 LOESS 平滑,得到本轮计算趋势项 T^{k+1} 。

在完成 n_i 轮内循环, T^k 与 S^k 收敛后,外循环将计算残差项 R ,并计算内循环步骤 2、3、6 中进行 LOESS 平滑时各数据点的权重 ρ 。外循环的计算公式如式(1)~式(4),median 表示求中位数。本研究中的 STL 分解算法使用 Python 的 statsmodels 模块实现。

$$R = Y - T - S \quad (1)$$

$$h = 6 \times \text{median}(|R|) \quad (2)$$

$$B(x) = \begin{cases} (1-x^2)^2 & (0 \leq x < 1) \\ 0 & (x \geq 1) \end{cases} \quad (3)$$

$$\rho = B\left(\frac{|R|}{h}\right) \quad (4)$$

其中: Y ——模板数据, mg/L;

h ——平滑参数, mg/L;

ρ ——进行 LOESS 平滑时各数据点的权重;

$B(x)$ ——权重计算函数。

1.2.3 合成数据构建

在对模板数据进行 STL 分解后,得到的残差项不含趋势性信息与周期性信息,且整体符合均值为 0 的正态分布。因此,在上述模板数据分解基础上,保留模板数据分解得到的 T 与 S ,使用均值为 0 的正态分布对 R 进行拟合,并对此拟合分布进行采样,生成高斯噪声,作为合成数据的 R 。最后将模板数据的 T 、 S 以及采样生成的合成数据的 R 进行加和,作为水库氯化物取水浓度合成数据。

1.2.4 其他方案参数设计

根据陈行水库实际运行情况,设定陈行水库初始水位为 6.0 m,以库内水位最低不低于 4.0 m,最高不高于 7.5 m 为控制条件,并设定取水泵站中的取水总流量为 0~100 m³/s,输水流量按实际情况设定为 12.73 m³/s,随机生成取水流量数据。考虑到取水泵站实际调节流量过程的要求,生成的取水流量数据满足最大连续取水时间与最大连续不取水时间均不超过 72 h,水泵开启的最小时间不低于 7 h,且保证库内水位不超出边界条件。此外,生成的模拟工况方案中考虑库内初始质量浓度为 50 mg/L 与 100 mg/L 的 2 种情形,风速风向采用模板时间段对

应的在线监测数据。

1.2.5 工况模拟汇总

综上所述,对于 1.2.1 节中选定的 3 组模板时间段,本研究依上述步骤分别各自构建了 5 组氯化物取水浓度合成数据(编号依次为 A~E)作为模拟工况方案的输入条件。对于每组氯化物取水浓度合成数据,本研究随机生成了 3 组取水流量模拟数据(编号依次为 a~c)作为模拟工况方案的取水流量输入条件。在此基础上,考虑上述条件在 1.2.4 节中选定的 2 种库内初始浓度下的情况,本研究共生成 90 个机理模型模拟工况方案,交由机理模型计算。

1.3 MIKE 21 机理模型构建

陈行水库的水深较小,水体在垂直方向上的混合较为均匀,因此,使用 MIKE 21 二维水动力水质模型对陈行水库进行机理模型构建。MIKE 21 的水动力模型部分使用有限体积法进行计算,其遵循二维不可压缩纳维-斯托克斯(Navier-Stokes)方程并同时满足流体静压假定和布辛涅斯克(Boussinesq)假定^[12]。水质模型部分采用 MIKE 21 的对流扩散模块模拟氯化物的分子扩散与对流传输,其通过对水质平衡方程进行离散化迭代求解。

本研究所使用的水动力与水质机理模型已经过详细率定^[13-14]。本研究使用的机理模型的地形数据由上海城投原水有限公司提供,模型详细参数如表 1 所示。模型水动力模块采用 2020 年陈行水库输水口实测水位资料以及 2021 年 5 月 7 日—2021 年 5 月 11 日、2021 年 12 月 26 日—2021 年 12 月 31 日的 2 次流场监测结果、实测风场和水库运行资料进行了率定和验证。结果表明,模型对 2 次流场监测结果的流速与流向模拟的平均相对误差(MRE)均不超过 0.3, NSE 分别为 0.61~0.94 和 0.53~0.79。模型的水质模块以 2021 年 3 月 1 日—3 月 15 日水库实测氯化物取水浓度作为边界条件,用水库实测氯化物输水浓度进行了率定。结果表明,模拟结果与实测数据的 MRE 为 0.085。在此基础上,2020 年 5 月—2021 年 4 月的陈行水库输水氯化物在线监测数据也被用于验证本研究使用的机理模型,验证表明 MRE 为 0.10、NSE 为 0.87。上述率定与验证结果显示,本研究使用的 MIKE 21 模型的水动力与水质模块均具有较高的模拟精度。

表 1 MIKE 21 模型参数

Tab. 1 The Parameters of Model MIKE 21

参数名称	参数值
涡粘函数	Smagorinsky 亚格子尺度模型
涡粘系数	0.28
曼宁数	13~58 m ^{1/3} /s
风摩擦系数	风速≤7 m/s 时,为 0.001 5 风速≥25 m/s 时,为 0.002 6 7 m/s<风速<25 m/s,根据上述两值进行线性内插
扩散系数	基于涡粘系数类比公式,涡粘系数比例因子参数取值为 1

1.4 LSTM 神经网络模型构建

1.4.1 数据准备与预处理

对于使用 MIKE 21 模型计算得到的机理模型模拟工况方案结果,从中提取库内水位变化数据以及氯化物输水浓度数据计算值,结合氯化物取水浓度合成数据、库内初始氯化物浓度、取水口泵站各水泵流量模拟数据、风向风速实测数据,作为 LSTM 神经网络模型训练阶段的输入值。本研究中 MIKE 21 模型与 LSTM 模型的数据流如图 3 所示。其中,由于风向数据与氯化物浓度、风速等数据不同,其角度值是出于人为规定,而不存在大小差异或自然顺序。因此,在本研究中将风向数据视为离散型变量,将风向分解为东、南、西、北、东南、西南、东北、西北共 8 个方向的独热(one-hot)编码^[15-16](如 22.5°~67.5°被视为西北风),也即分解为 8 列的矩阵,样本的每个时间点数据将对应方向的矩阵值设置为 1,其余设置为 0。为使得神经网络能够更好地提取输入数据的特征,对所有方案的除风向数据外的其他神经网络输入数据整体进行标准化,以确保输入数据整体处于[0, 1]。

1.4.2 LSTM 基本原理

RNN 是一类专门设计用于学习序列数据的神经网络模型,在实际应用中其通常被用于学习时间序列数据的特征^[17]。Hochreiter 等^[18]提出的 LSTM 神经网络是对传统 RNN 的有效改进,其通过在已有隐藏层基础上增加记忆元结构以存储过去的长期记忆,并使用遗忘门控制其保留程度。该设计实现了神经网络对输入信息的长期保留,从而解决了 RNN 由于梯度消失而难以捕捉到长距离依赖关系的问题^[19]。LSTM 神经网络的隐藏层结构如图 4 所示。其中 H_{t-1} 与 H_t 分别为上一次与本次的隐状态输出

(其中 H_t 同时为隐藏层的输出), C_{t-1} 和 C_t 分别为上一次与本次记忆元输出, X_t 为本次隐藏层输入, $\tilde{C}$ 为候选记忆元, I 为输入门, F 为遗忘门, O 为输出门。

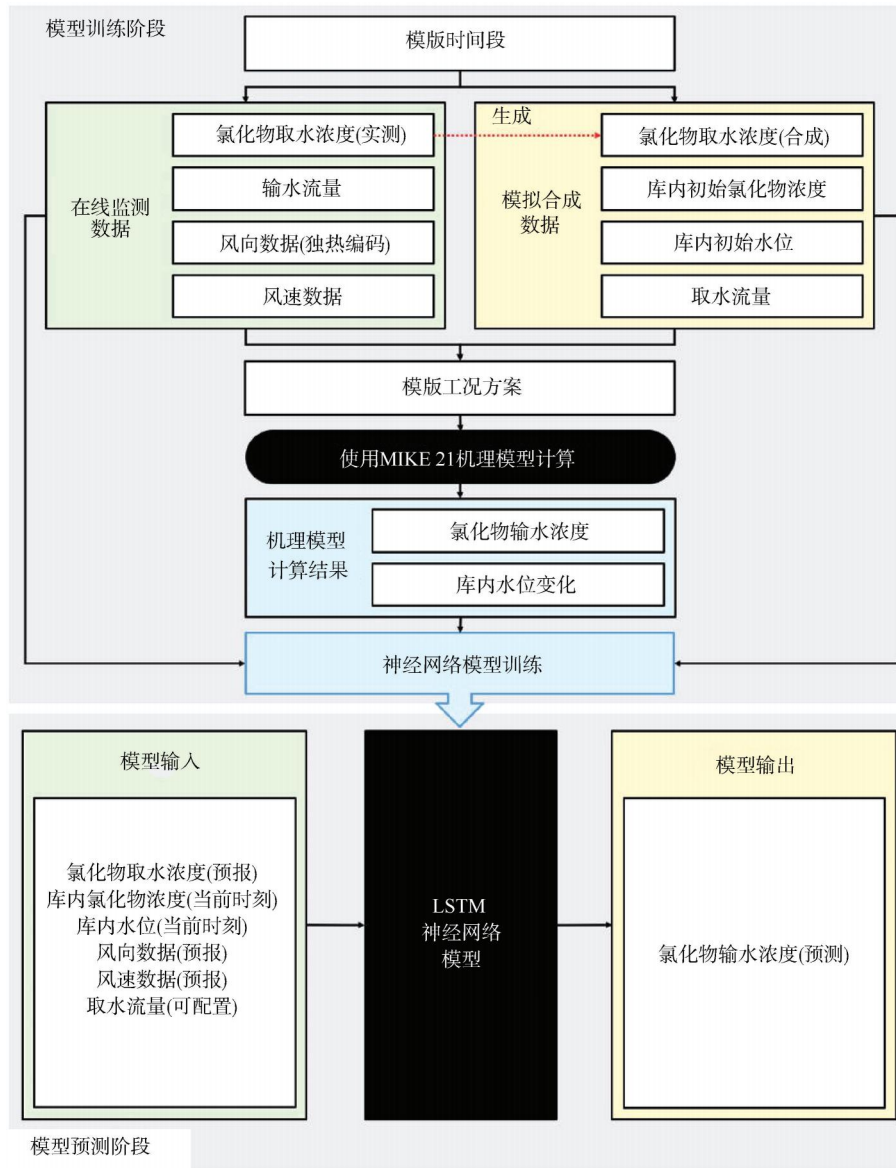


图3 MIKE 21 与 LSTM 模型数据流

Fig. 3 Data Flow of MIKE 21 and LSTM Models

由图 4 可知,候选记忆元 $\tilde{C}$ 与传统 RNN 的隐藏层类似(但激活函数不同),输入门 I_t 负责决定 $\tilde{C}_t$ 的输出的采用程度, F 负责决定 C_{t-1} 的保留程度, O 负责决定 H_t 。此外,图 4 中 $\odot$ 表示计算矩阵的 Hadamard 积, $\tanh$ 为双曲正切函数, σ 为 Sigmoid 函数,如式(5)~式(12)。

$$I_t = \sigma([X_t, H_{t-1}] \times W_I + b_I) \quad (5)$$

$$F_t = \sigma([X_t, H_{t-1}] \times W_F + b_F) \quad (6)$$

$$O_t = \sigma([X_t, H_{t-1}] \times W_O + b_O) \quad (7)$$

$$\tilde{C} = \tanh([X_t, H_{t-1}] \times W_{\tilde{C}} + b_{\tilde{C}}) \quad (8)$$

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t \quad (9)$$

$$H_t = O_t \odot \tanh(C_t) \quad (10)$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (11)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (12)$$

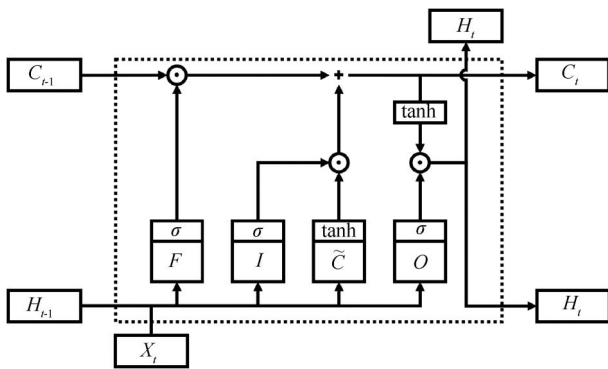


图4 LSTM 隐藏层结构

Fig. 4 Structure of Hidden Layer in LSTM

其中: I_t 、 F_t 、 O_t 与 $\tilde{C}_t$ ——依次表示输入门、遗忘门、输出门、候选神经元的输出值;
 W_I 、 W_F 、 W_O 与 $W_{\tilde{C}}$ ——依次表示输入门、遗忘门、输出门、候选神经元的权重;
 b_I 、 b_F 、 b_O 与 $b_{\tilde{C}}$ ——依次表示输入门、遗忘门、输出门、候选神经元的偏置项。

1.4.3 模型训练

对于所有模拟工况方案的计算结果,按约 7:1:2 的比例将所有结果样本划分为训练集、检验集与测试集。使用 MSE 作为损失函数,采用基于时间的反向传播(back propagation through time, BPTT)对神经网络参数进行更新和训练。为保证神经网络的训练效果,在模型训练前使用 Kaiming 参数初始化^[20],并在训练过程中使用参数 $\theta = 1$ 的梯度裁剪加速训练并避免梯度爆炸^[21]。神经网络训练的优化算法选择 Kingma 等^[22]提出的自适应矩估计(adaptive moment estimation, Adam)优化算法,该优化算法简单高效,并且在许多情况下具有良好的应用效果^[23]。模型采用 1 个输入层,1 个隐藏层与 1 个输出层的结构,并使用 ReLU 函数作为激活函数。

1.4.4 模型评估

采用均方根误差(root-mean-square error, RMSE)、MRE 与 NSE 对训练完成的 LSTM 模型的效果进行评估。其中, RMSE 与 MRE 越接近于 0, NSE 越接近于 1, 说明模型的效果越好。相应计算如式(13)~式(15)。每个样本的上述评估参数各自单独计算。

$$R = \sqrt{\frac{1}{n} \sum_{t=1}^n (\hat{y} - y)^2} \quad (13)$$

$$M = \frac{1}{n} \sum_{t=1}^n \left| \frac{\hat{y} - y}{y} \right| \quad (14)$$

$$N = 1 - \frac{\sum_{t=1}^n (\hat{y} - y)^2}{\sum_{t=1}^n (y - \bar{y})^2} \quad (15)$$

其中: R ——RMSE 值, mg/L;

M ——MRE 值;

N ——NSE 值;

n ——模型预测的步数;

y ——对应时间步氯化物输水质量浓度的机理模型计算值, mg/L;

$\hat{y}$ ——对应时间步氯化物 LSTM 神经网络预测值, mg/L;

$\bar{y}$ ——单个样本所有时间步氯化物输水质量浓度的机理模型计算值的平均值, mg/L。

1.4.5 模型预测

在模型 LSTM 模型训练完成后,模型可不再依赖 MIKE 21 机理模型而进行独立地预测和使用。如图 3 所示,在预测时,模型的氯化物取水浓度、风向风速数据均来自未来待预测时段的预报值,库内氯化物浓度和库内水位均为当前时刻的实测值,而取水水泵在未来待预测时段的流量数据可根据模型使用者的需要自行设置。

2 结果与讨论

2.1 STL 分解与模拟工况方案生成

为使得生成的水库氯化物取水浓度合成数据能够尽可能包含在线监测数据的趋势性与周期性特征,分析了参数 n_p 的不同取值对 STL 分解结果的影响,并以分解后趋势项一阶差分的标准差 $\sigma[\Delta T_{(1)}]$ 以及残差项的标准差 $\sigma(R)$ 作为分解的效果评估指标。探索了 n_p 取值分别为 6、12、24、48、96 的情况。此外,为使得 STL 分解结果满足 Cleveland 等^[11]提出的矩阵特征值(Eigenvalue)标准,选定 n_p 后的其他参数根据 Cleveland 等^[11]的建议值决定。

不同 n_p 取值下的分解效果评估指标如图 5 所示。由图 5 可知,当参数 n_p 的取值增大时,各模板数据的 $\sigma[\Delta T_{(1)}]$ 逐渐下降且趋于平缓,且 n_p 较小时, $\sigma[\Delta T_{(1)}]$ 显著较高。一阶差分数据由相邻 2 项数据计算差值而来,其可以部分反映数据的平滑程度。因此,这表明当 n_p 取值较小时, STL 分解的趋

势项不够平滑。这可能是由于过小的参数 n_p 使得部分数据波动信息被错误地纳入了趋势项。而与之相反,模板数据的 $\sigma(R)$ 随参数 n_p 的取值增大而显著提升,这表明过大的参数 n_p 的取值会降低 STL 分解对模板数据中趋势性与周期性特征的提取能力。根据图中所示,当 $n_p = 12$ 时,STL 分解的趋势项与残差项分解效果均较好。这可能是因为 STL 分解中的 n_p 参数作为循环子序列的分组间隔,其部分体现了时序数据的周期性特征。而根据黄惠明等^[24]对长江口潮汐特征的分析,长江口潮汐具有良好的规律,即 1 个太阳日中会出现 2 次高潮与低潮。因此, $n_p = 12$ 的参数选择与长江口的潮汐周期相近,是使得该参数下 STL 分解效果较好的可能原因。本研究中的后续模拟工况方案均选取 $n_p = 12$ 的参数,其他参数根据上述 Cleveland 等^[11]的建议选用 $n_s = 7$ 、 $n_1 = 13$ 、 $n_1 = 23$ 。

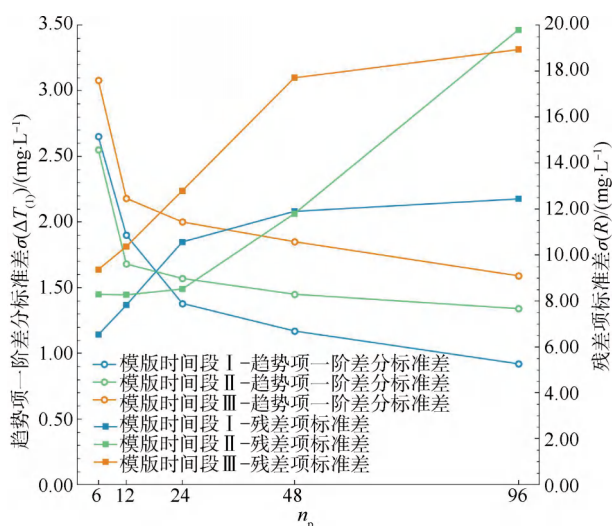


图5 模板数据不同参数下 STL 分解效果评估

Fig. 5 Evaluation of STL Decomposition Effect of Template Data under Different Parameters

对于本研究中选定的 3 段模板时间段,分别使用上述参数对模板数据进行 STL 分解,分解结果如图 6 所示。由图 6 可知,分解后得到的残差项总体均匀分布在 0 的两侧且服从正态分布。对于氯化物输水浓度数据,此处分解得到的残差项可能来自在线监测仪器自身的随机性误差,也有可能来自在线监测仪器附近水体受风场、长江上游来水的综合影响。对于此类偶发性不规则数据波动,根据中心极限定理并出于对测量误差对称性与无偏性的考虑,本研究认为其服从均值为 0 的正态分布,并进一步

基于此生成的高斯噪声构建合成数据。由于分解后残差项在不同时间段的散布程度不同,为使得构建的合成数据也保留该特征,本研究以 $n_p = 12$ 个数据为间隔对模板数据分组,每组各自计算残差项的标准差,并根据此标准差基于正态分布生成高斯噪声,作为该段时间合成数据的残差项。将所有组合成数据残差项拼接后,将其与模板数据的趋势项、季节项进行加和,构成合成数据,如图 6 所示。由图 6 可知,合成数据在保留了模板数据的趋势项与周期性特征基础上,产生了额外的波动。

2.2 LSTM 神经网络在不同隐藏层神经元数量与预测时间步长下的预测效果

为对比在不同的 LSTM 隐藏层神经元数量以及预测时间步长下模型对水库氯化物输水浓度的预测能力,分别选取 16、32、64、128、256 作为隐藏层神经元个数,选取 1、2、3、6、12、24、48 h 作为预测时间步长,基于相同的训练集、验证集与测试集样本划分,对 LSTM 模型进行训练。同时,计算了各模型的 RMSE、MRE 与 NSE 作为模型的评估参数,如图 7 所示。

由图 7 可知,不同的 LSTM 隐藏层神经元数量以及预测时间步长对模型的预测能力均有较为显著的影响。当 LSTM 隐藏层神经元数量上升,模型的预测能力总体上呈现出先上升再下降的趋势。如当预测时间步长为 6 h 时,当 LSTM 隐藏层神经元数量为 16 与 128 时,训练集的 RMSE 较高(分别为 0.12 mg/L 与 0.11 mg/L),当隐藏层神经元数量为 32 与 64 时较低(为 0.09 mg/L);当预测时间步长为 24 h 时,在隐藏层神经元从 16 提升至 128 的过程中, RMSE 不断减小(训练集 RMSE 从 0.16 mg/L 逐步降低至 0.11 mg/L,验证集 RMSE 从 0.17 mg/L 逐步降低至 0.13 mg/L,测试集从 0.19 mg/L 逐步降低至 0.13 mg/L)、MRE 不断减小(训练集 MRE 从 0.05 逐步降低至 0.03,验证集 MRE 从 0.06 逐步降低至 0.04,测试集从 0.06 逐步降低至 0.04),同时 NSE 不断增大(训练集 NSE 从 0.95 逐步提升至 0.98,验证集 MRE 从 0.93 逐步提升至 0.96,测试集 MRE 从 0.93 逐步提升至 0.96),但从 128 提升至 256 时出现了相反的趋势。这可能是因为,当模型的 LSTM 隐藏层神经元数量较低时,模型的学习能力不足以掌握水库复杂的水动力与水质扩散过程,因而难以实现精度较高的氯化物浓度预测;而当

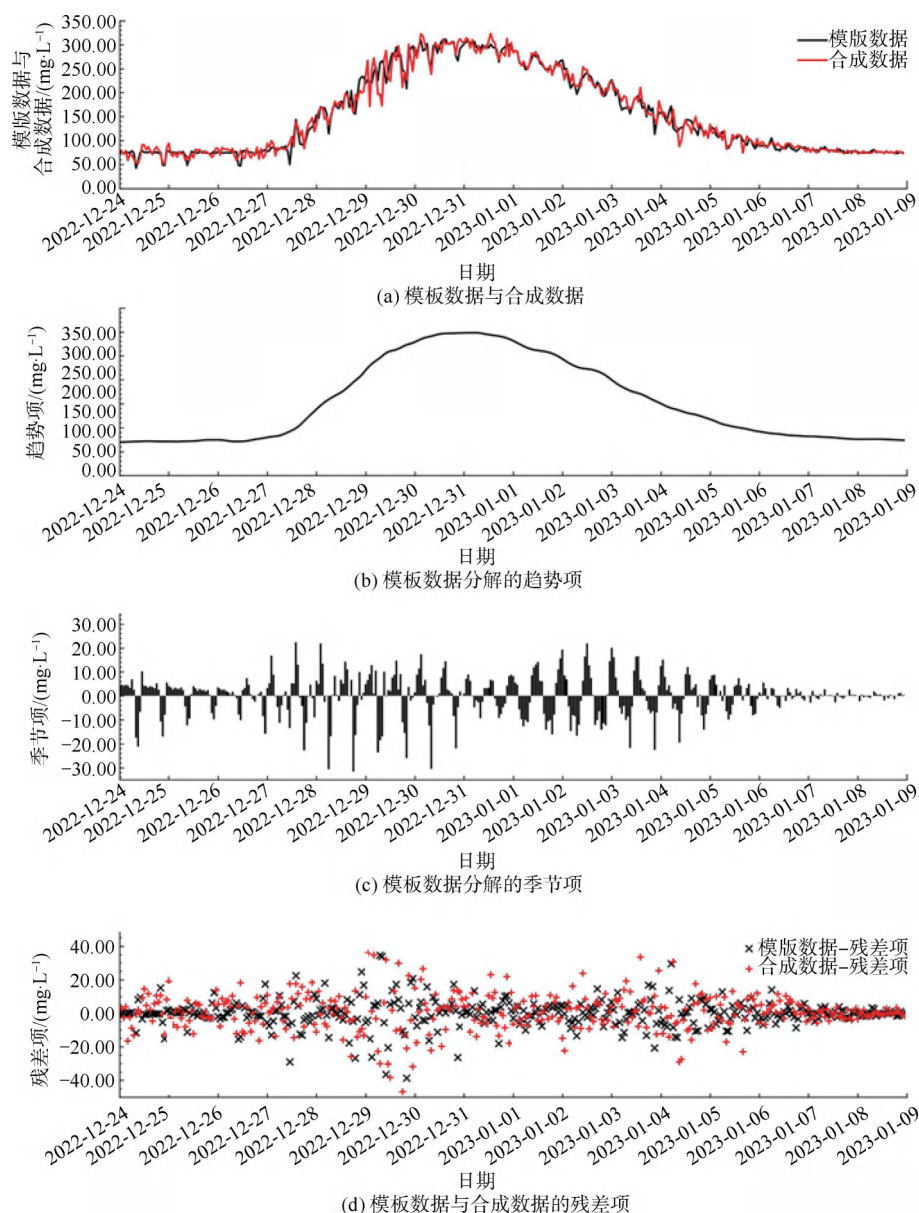


图6 STL分解与合成数据构建(以模板时间段Ⅱ为例)

Fig. 6 STL Decomposition and Synthetic Data Generation (Example of Template Period II)

隐藏层神经元数量过高,模型神经元参数的训练需要依赖于更大的训练集样本规模。因此,预测能力出现了下降。

当模型的预测时间步长上升时,模型的预测能力出现持续下降趋势。并且由图7可知,预测时间步长从24 h提升至48 h时的预测能力下降尤为显著,其体现在RMSE与MRE的较前者的显著上升以及NSE的显著下降。其原因可能有2点:一是当预测时间步长过短时,水库本身的氯化物浓度波动不大,此时的模型虽然预测能力较强,但模型的实际应

用价值较差,因为过短的预测时间步长难以为水库决策者预留足够的应对咸潮入侵的反应时间;二是当预测时间步长过长时,每一轮训练需要传入的样本步长增加,其等价于可用于训练的样本数据量降低,模型的训练需要更大的训练集样本规模与时间长度。考虑到机理模型的计算时间往往较长,扩大模型训练数据集的代价极大,综合考虑不同的隐藏层神经元数量以及预测时间步长下的LSTM神经网络模型的效果以及模型的实际应用价值,本研究认为隐藏层神经元数量为128以及预测时间步长为

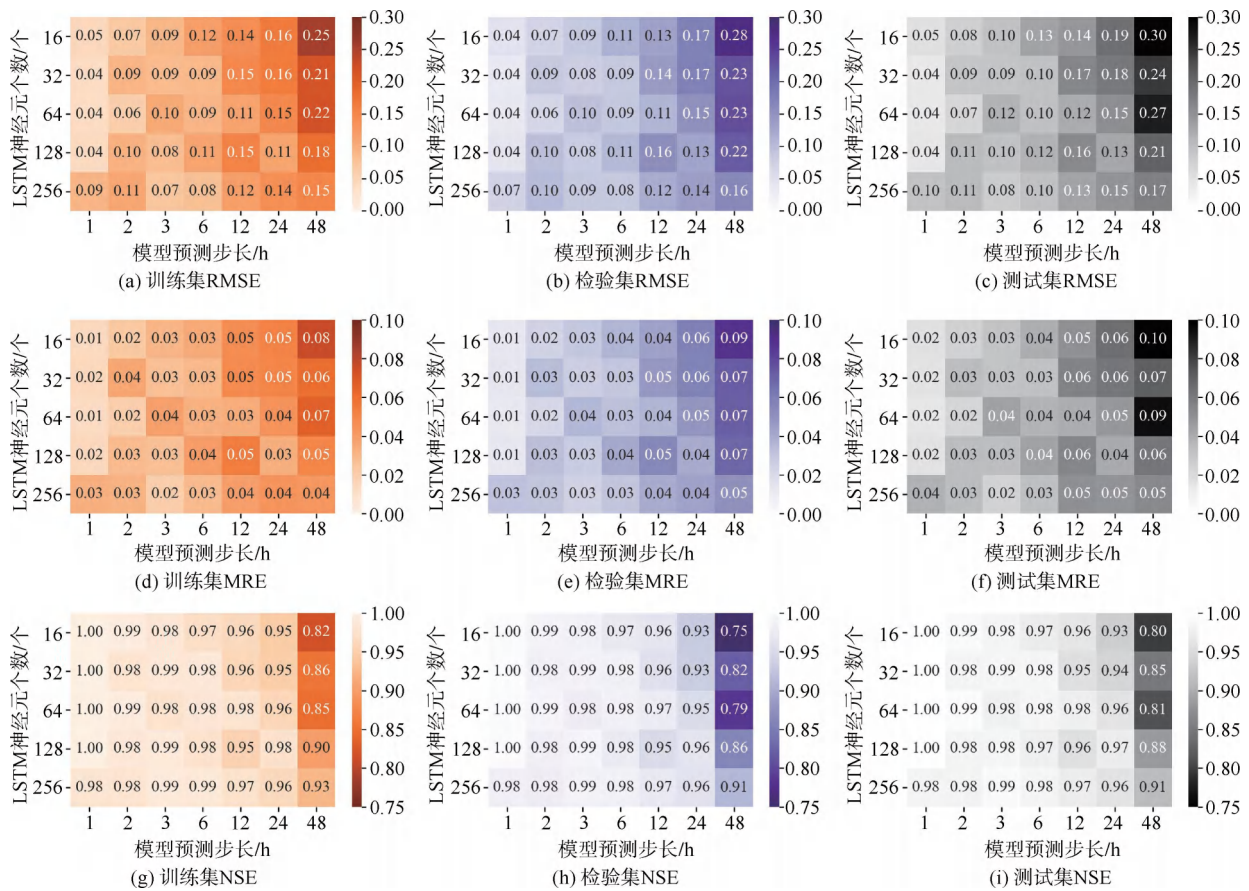


图7 不同 LSTM 隐藏层神经元个数与预测时间步长下的 RMSE、MRE 与 NSE

Fig. 7 RMSE, MRE and NSE under Different Numbers of LSTM Hidden Layer Neurons and Prediction Time Steps

24 h 是较好的选择。

2.3 不同结构神经网络模型预测能力对比分析

为进一步探索不同结构的神经网络对水库盐度的预测效果,本研究选取了反向传播神经网络(BPNN)、一维卷积神经网络(1dCNN)、RNN 与门控

神经网络(GRU)作为经典的神经网络结构,采用2.2节中选取的隐藏层神经元数量与预测步长,基于同样的数据集划分进行模型训练,并将模型训练结果与使用 LSTM 建立的模型进行对比。各模型的训练结果参数评估如表2所示。

表2 不同神经网络结构训练效果对比

Tab. 2 Effects of Different Neural Network Structures

神经网络 结构名称	训练集(63组方案)			验证集(7组方案)			测试集(20组方案)		
	RMSE	MRE	NSE	RMSE	MRE	NSE	RMSE	MRE	NSE
BPNN	0.22	0.07	0.90	0.21	0.07	0.92	0.25	0.08	0.91
1dCNN	0.20	0.06	0.92	0.18	0.06	0.95	0.23	0.08	0.92
RNN	0.19	0.06	0.93	0.20	0.06	0.93	0.23	0.08	0.92
GRU	0.18	0.06	0.93	0.17	0.05	0.95	0.21	0.06	0.93
LSTM	0.11	0.03	0.98	0.13	0.04	0.96	0.13	0.04	0.96

由表2可知,对于此5种不同结构的神经网络在训练集、验证集与测试集上的表现,使用 RMSE、MRE 与 NSE 进行表征,数据整体显示出相当一致的趋势。首先从总体上看,3种循环神经网络

(RNN、GRU 与 LSTM)的预测能力好于2种非循环神经网络(BPNN 与 1dCNN),这说明对于水库氯化物输水浓度这类复杂环境问题的模拟和预测,循环神经网络基于从过往隐藏层输出获取信息的设计能

够更好地从数据的时序特征中获取重要信息。在 2 种非循环神经网络 (BPNN 与 1dCNN) 中, 1dCNN 的预测能力好于 BPNN, 这可能是因为 BPNN 的隐藏层仅考虑了当前时刻的输入值, 而 1dCNN 虽然未从过往隐藏层输出获取信息, 但其卷积核在数据时间维度上的滑动也获取了输入数据的时间局部特征^[25], 使其相较于 BPNN 有较好的预测效果。在 3 种循环神经网络中, RNN 与 GRU 的测试集预测效果相近, 而 LSTM 的测试集预测效果显著好于 RNN 与 GRU, 其体现在 LSTM 的测试集 RMSE (0.13 mg/L) 显著低于 RNN 与 GRU 的测试集 RMSE (分别为 0.23 mg/L 与 0.21 mg/L), LSTM 的 NSE (0.96) 显著高于 RNN 与 GRU 的 NSE (分别为 0.93 与 0.92)。这意味着在相同的隐藏层神经元数量下,

对于相同步长的水库氯化物输水浓度预测任务, LSTM 具有更好的表现效果。考虑到已有研究^[13]表明陈行水库内氯化物浓度分布与库区水龄、取水流量之间存在复杂的非线性关系, 上述试验结果验证了 LSTM 模型在学习此类关系时相较于其他神经网络结构所具有的更大优势。

2.4 神经网络模型预测精度验证与求解速度分析

为进一步验证基于本研究方法训练构建得到的 LSTM 神经网络模型的可信度与实际运用于陈行水库的氯化物浓度预测的可行性, 本研究根据 1.4.5 节中所述, 选取 2023 年 3 月 27 日—2023 年 4 月 5 日的陈行水库对应时刻实测数据与预报数据作为模型输入, 氯化物输水浓度作为输出, 对建立的 LSTM 神经网络模型进行了验证。验证效果如图 8 所示。

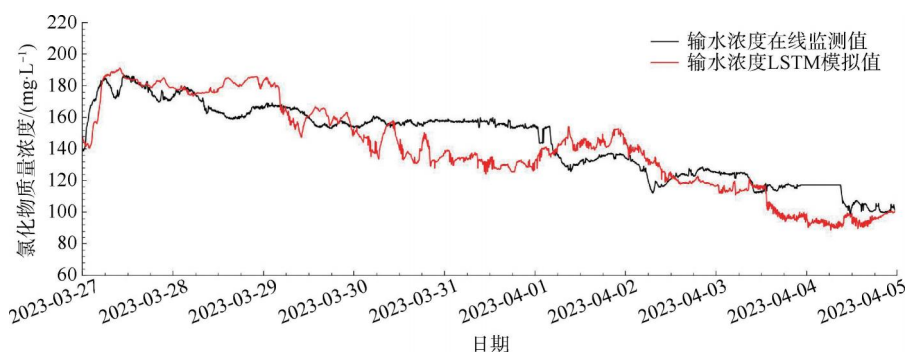


图 8 输水浓度在线监测值与 LSTM 神经网络模拟值对比

Fig. 8 Comparison between Online Monitoring Values and Simulated Values of LSTM

由图 8 可知, 在此时间段内, 氯化物输水质量浓度在线监测数据存在从约 140 mg/L 上升至约 190 mg/L, 而后逐步下降至约 100 mg/L 的过程。LSTM 神经网络模型对此浓度变化过程趋势的模拟总体上与在线监测数据较为一致, 其中对咸潮峰值质量浓度的模拟精度可达到 10 mg/L 以内, 对全过程模拟的最大质量浓度误差在 20 mg/L 以内。使用评估参数对 LSTM 神经网络模型的全过程模拟进行表征, 发现 RMSE 可达 0.29 mg/L、MRE 可达 0.09、NSE 可达 0.58, 这表明使用 MIKE 21 模拟工况方案计算结果建立的 LSTM 模型保有了较高的预测模拟精度, 一定程度上具备了替代机理模型进行咸潮入侵时的应对方案求解的能力。

除计算精度外, 本研究建立的 LSTM 神经网络模型较之机理模型在计算速度上有极大的提升, 对算力的需求极大降低。本研究中使用的工作站硬件配置下, 使用 MIKE21 求解 1.2.5 节中建立的 90 个

机理模型模拟工况方案时, 各机理模型的平均计算时间约为 45 min, 且由于 MIKE 21 的算力要求较高, 计算时完全占用中央处理器 (CPU) 全部核心, 无法通过同时并行计算多个方案实现平均计算速度的提升。而使用本研究建立的 LSTM 神经网络模型对 90 个机理模型模拟工况计算得到的训练集、检验集与测试集样本进行预测计算时, 在基于 Python 的 multiprocessing 模块建立核心进程数为 15 的进程池进行并行计算的前提下, 可以实现在约 10 s 内完成全部计算。这意味着在相同的时间内, 使用本研究建立的 LSTM 神经网络模型能够进行更多次数的试算, 这对咸潮入侵时的辅助决策能够提供更高的价值。

3 结论

本研究探究了使用神经网络模型替代湖泊水库非线性水文水质过程机理模型, 实现快速计算在咸潮入侵期间不同取水条件下的输水水质的可行性,

提出了一种耦合了 STL 分解、基于高斯噪声的数据增强、MIKE 21 机理模型与 LSTM 神经网络模型的 STL-MIKE-LSTM 联合建模方法,并将该方法运用于陈行水库。通过探究不同参数下 STL 分解的时序特征提取效果,发现选择 $n_p = 12$ 使 STL 分解的效果较好。通过对比不同隐藏层神经元个数与预测步长下 LSTM 模型的预测效果,发现预测效果随神经元个数增加先升后降,随预测步长增加持续下降,选取神经元个数为 128 与预测步长为 24 h 的总体效果较好。通过对比不同结构神经网络的预测能力,发现 LSTM 在预测集预测效果最佳 ($RMSE = 0.13 \text{ mg/L}$, $MRE = 0.04$, $NSE = 0.96$)。通过使用实际实测数据与预报数据验证 LSTM 模型,发现其对预测氯化物输水浓度具备较高精度 ($RMSE = 0.29 \text{ mg/L}$, $MRE = 0.09$, $NSE = 0.58$),且所需算力与时间远低于 MIKE 21 机理模型。综上,本研究提出的湖泊水库氯化物输水浓度预测方法经验证同时具备较高计算精度与速度,一定程度具备替代机理模型为水库管理者应对咸潮入侵提供快速决策支持的能力。考虑到受模板时间段及训练样本所限,本研究提出的联合建模方法建立的 LSTM 神经网络模型仅适用于中高取水浓度条件下的输水浓度快速预测,而不适用于日常低取水浓度条件或极端超高取水浓度条件。未来的研究可进一步提升模型的预测精度与鲁棒性,并考虑基于本研究成果探索决策的自动生成策略。

参考文献

- [1] 王玉琦,李铖,刘安琪,等. 2022 年长江口夏季咸潮入侵及影响机制研究[J]. 人民长江, 2023, 54(4): 7-14.
WANG Y Q, LI C, LIU A Q, et al. Dynamic mechanism of saltwater intrusion at Yangtze River estuary in summer, 2022 [J]. Yangtze River, 2023, 54(4): 7-14.
- [2] 乔红杰,刘大伟,闻卫东,等. 2022 年长江口北支咸水倒灌传播特征分析[J]. 人民长江, 2023, 54(2): 63-69.
QIAO H J, LIU D W, WEN W D, et al. Diffusion characteristics of saltwater intrusion in north branch of Yangtze River Estuary in 2022 [J]. Yangtze River, 2023, 54(2): 63-69.
- [3] NORIO N, MIYAZAKI K. The analysis of saltwater intrusion through Komesu underground dam and water quality management for salinity[J]. Paddy and Water Environment, 2009, 7: 71-82. DOI: 10.1007/s10333-009-0154-1.
- [4] 许婷. 丹麦 MIKE21 模型概述及应用实例[J]. 水利科技与经济, 2010, 16(8): 867-869.
XU T. Calculation principle and application example of a two-dimensional flow model-MIKE21 HD [J]. Water Conservancy Science and Technology and Economy, 2010, 16(8): 867-869.
- [5] 夏妍,窦鹏,楼春华,等. 不同植被覆盖情景下景观水体水质改善效果模拟研究[J]. 环境科学学报, 2022, 42(1): 248-258.
XIA Y, DOU P, LOU C H, et al. Simulation study on the effect of water quality improvement of landscape water body under different vegetation coverage scenarios [J]. Acta Scientiae Circumstantiae, 2022, 42(1): 248-258.
- [6] PARVAZE S, JAIN M K, ALLAIE S P. Integrated hydrologic and hydraulic flood modelling for a scarcely gauged inter-montane basin: A case study of Jhelum Basin in Kashmir Valley [J]. Sadhana-Academy Proceedings in Engineering Sciences, 2023, 48(1). DOI: 10.1007/s12046-022-02072-1.
- [7] 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.
- [8] LIANG Z Y, ZOU R, CHEN X, et al. Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach [J]. Journal of Hydrology, 2020, 581: 124432. DOI: 10.1016/j.jhydrol.2019.124432.
- [9] ZHENG L, WANG H P, LIU C, et al. Prediction of harmful algal blooms in large water bodies using the combined EFDC and LSTM models [J]. Journal of Environmental Management, 2021, 295: 113060. DOI: 10.1016/j.jenvman.2021.113060.
- [10] YOO J, KANG S. Class-adaptive data augmentation for image classification [J]. IEEE Access, 2023, 11: 26393-26402. DOI: 10.1109/ACCESS.2023.3258179.
- [11] CLEVELAND R B, CLEVELAND W S, TERPENNING I, et al. STL: A seasonal-trend decomposition procedure based on Loess [J]. Journal of Official Statistics, 1990, 6(1): 3-3.
- [12] WARREN I R, BACH H K. MIKE 21: A modelling system for estuaries, coastal waters and seas [J]. Environmental Software, 1992, 7(4): 229-240.
- [13] 陈思绪,崔婧嫒,黄熠丽,等. 基于水龄的陈行水库流场特征分析[J]. 净水技术, 2023, 42(3): 60-66, 119.
CHEN S X, CUI J Y, HUANG Y L, et al. Analysis of flow field characteristics for Chenhang Reservoir based on the water age [J]. Water Purification Technology, 2023, 42(3): 60-66, 119.
- [14] 陈思绪. 湖库水质对水龄的响应研究——以华东地区某饮用水水库为例[D]. 上海: 同济大学, 2023.
CHEN S X. Response of water quality to water age in lakes and reservoirs: A case study of a drinking water reservoir in east China [D]. Shanghai: Tongji University, 2023.

(下转第 186 页)

- 放核算及其特征[J]. 中国给水排水, 2022, 38(21): 39-44.
- QIAN X Y, HU J, LI D, et al. Calculation and characteristics of greenhouse gas emission from urban wastewater treatment plants in shanghai [J]. China Water & Wastewater, 2022, 38(21): 39-44.
- [3] 张惠. 城镇污水处理行业低碳技术研究现状与发展趋势分析[J]. 净水技术, 2024, 43 (7): 1-9.
- ZHANG H. Analysis on research status and development trend of low-carbon treatment and disposal technologies for urban sewage and sludge [J]. Water Purification Technology, 2024, 43(7): 1-9.
- [4] 方能峰. 类IV类出水水质污水处理厂的设计及运行效果分析[D]. 合肥: 合肥工业大学, 2020.
- FANG N F. Design and operation performance analysis of wastewater treatment plant with water-quality near IV class effluent discharging standard [D]. Hefei: Hefei University of Technology, 2020.
- [5] 张羽就,席佳锐,陈玲,等. 中国城镇污水处理厂能耗统计与基准分析[J]. 中国给水排水,2021,37(8):8-17.
- ZHANG Y J, XI J R, CHEN L, et al. Energy consumption statistics and benchmarking analysis of urban wastewater treatment plants (WWTPs) in China [J]. China Water & Wastewater, 2021, 37(8): 8-17.
-
- (上接第 68 页)
- [15] CERDA P, VAROQUAUX G, KÉGL B. Similarity encoding for learning with dirty categorical variables[J]. Machine Learning, 2018, 107(8): 1477-1494.
- [16] KRISTIANI E, LIN H, LIN J R, et al. Short-term prediction of PM_{2.5} using LSTM deep learning methods [J]. Sustainability, 2022, 14: 2068. DOI: 10.3390/su14042068.
- [17] 张弛, 郭媛, 黎明. 人工神经网络模型发展及应用综述[J]. 计算机工程与应用, 2021, 57(11): 57-69.
- ZHANG C, GUO Y, LI M. Review of development and application of artificial neural network models [J]. Computer Engineering and Application, 2021, 57(11): 57-69.
- [18] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.
- [19] AHMED D M, HASSAN M M, MSTAFA R J. A review on deep sequential models for forecasting time series data[J]. Applied Computational Intelligence and Soft Computing, 2022. DOI: 10.1155/2022/6596397.
- [20] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[J]. arXiv, 2015: 01852. DOI: 10.48550/arXiv.1502.01852.
- [21] ZHANG J Z, HE T X, SRA S, et al. Why gradient clipping accelerates training: A theoretical justification for adaptivity[C]. Ethiopia: International Conference on Learning Representations, 2020.
- [22] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv, 2017: 6980. DOI: 10.48550/arXiv.1412.6980.
- [23] SONG C Y, ZHANG H P. Study on turbidity prediction method of reservoirs based on long short term memory neural network [J]. Ecological Modelling, 2020, 432: 109210. DOI: 10.1016/j.ecolmodel.2020.109210.
- [24] 黄惠明, 姚佳辉, 王义刚, 等. 长江口北槽枯季混合与层化特征分析[J]. 河北工程大学学报, 2020, 37(1): 66-74.
- HUANG H M, YAO J H, WANG Y G, et al. Analysis of mixing and stratification in dry season of the north passage of the Yangtze River estuary[J]. Journal of Hebei University of Engineering, 2020, 37(1): 66-74.
- [25] ALZUBAIDI L, ZHANG J, HUMAIDI A J, et al. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions[J]. Journal of Big Data, 2021, 8 (1): 53-53.