

DOI:10.19853/j.zgjsps.1000-4602.2025.19.023

基于深度学习的污水检查井液位峰值预测方法

杨利刚¹, 郭 帅¹, 申屠华斌²

(1. 合肥工业大学 土木与水利工程学院, 安徽 合肥 230009; 2. 中国电建集团华东勘测设计研究院有限公司, 浙江 杭州 311122)

摘 要: 污水检查井液位峰值预测可为提前发现排水管网潜在问题并采取预防措施以避免污水溢流提供重要的依据。相对于传统时序预测方法,深度学习具有更强大的拟合能力和特征学习能力。基于W市某区布设的4个在线液位计2023年1月—12月的每小时液位峰值数据,分别搭建基于随机森林(RF)、BP神经网络和长短期记忆(LSTM)神经网络3种方法的液位峰值预测模型,并采用均方根误差(RMSE)、平均绝对误差(MAE)、决定系数(R^2)等指标评价不同预测模型的性能表现。结果显示,基于同一数据集所建立的3种预测模型的 R^2 仅存在细微差异且均接近于1,3种预测方法均表现出良好的拟合程度;LSTM神经网络模型的MAE和RMSE值均最小,表明在预测误差和稳定性方面该模型优于另两种模型,LSTM神经网络模型具有更优的泛化能力。

关键词: 污水检查井; 液位峰值预测; 深度学习; 长短期记忆(LSTM)神经网络; 时间序列预测模型

中图分类号: TU992 文献标识码: A 文章编号: 1000-4602(2025)19-0163-06

Prediction Method of Peak Liquid Level in Sewage Manhole Based on Deep Learning

YANG Li-gang¹, GUO Shuai¹, SHENTU Hua-bin²

(1. School of Civil and Hydraulic Engineering, Hefei University of Technology, Hefei 230009, China; 2. Power China Huadong Engineering Corporation Limited, Hangzhou 311122, China)

Abstract: The prediction of peak liquid levels in sewage manholes can serve as a critical reference for the early identification of potential issues within drainage networks, enabling proactive measures to prevent sewage overflow. Compared with traditional time series prediction methods, deep learning demonstrates superior capabilities in both function fitting and feature learning. Based on the hourly peak liquid level data collected from four on-line liquid level meters installed in a district of city W between January and December 2023, peak liquid level prediction models were developed using three different methodologies: random forest (RF), back propagation (BP) neural network, and long short-term memory (LSTM) neural network. The performance of these models was evaluated using key evaluation metrics, including root mean square error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). The R^2 values of the three prediction models, established based on the same dataset, were very similar and all approach 1, indicating a high degree of model fit for all three predictive methods. The MAE and RMSE values of the LSTM neural network model were the lowest among the three models,

基金项目: 安徽省重点研发计划项目(202104i07020012); 国家重点研发计划项目(2024YFC3810803)

通信作者: 郭帅 E-mail: guoshuai@hfut.edu.cn

indicating that it outperformed the other two in terms of prediction accuracy and stability. Furthermore, the LSTM model demonstrated superior generalization capability.

Key words: sewage manhole; peak liquid level prediction; deep learning; long short-term memory (LSTM) neural network; time series prediction model

当前,我国部分城市仍面临排水管网配套建设滞后、管网老化严重等问题,导致污水检查井溢流事件时有发生,对环境和公共健康构成严重威胁^[1-2]。准确预测污水检查井的液位峰值有助于提前发现溢流风险,从而采取预防措施,减少溢流、内涝等事件的发生。现有的城市排水管网液位监测与预警系统主要依赖于设定液位计报警阈值,其预警功能相对单一,缺乏对未来变化的预见性和主动性^[3]。开发基于深度学习算法的检查井液位时间序列预测模型,可对历史液位数据动态特性进行系统化学习,并在不断更新的数据中进行自适应调整,从而实现对液位峰值发生情况的准确预测,对主动预测预防溢流事件的发生具有重要意义。

常见的时间序列预测模型包括自回归移动平均模型、支持向量机(SVM)、随机森林(RF)、循环神经网络(RNN)、卷积神经网络(CNN)、BP神经网络、长短期记忆(LSTM)神经网络等,其中BP神经网络、LSTM神经网络等深度学习模型具有强大的拟合能力和特征学习能力,在水资源相关领域得到了广泛应用^[4-7],但对污水检查井液位峰值的预测却鲜有研究。为此,利用W市某区布设的4个污水检查井液位计2023年1月—12月的监测数据,分别构建随机森林、BP神经网络、LSTM神经网络3种时间序列预测模型,对污水检查井未来液位峰值变化进行预测,并以均方根误差(RMSE)、平均绝对误差(MAE)、决定系数(R^2)作为评价指标来选择最优模型,实现对未来液位峰值的精准预测。

1 研究方法及数据处理

1.1 3种时间序列预测方法

随机森林是一种集成学习算法,通过构建多个决策树组成“森林”来进行分类或回归。在构建每个决策树时,随机森林采用随机抽取的数据子集和随机选择的特征,以增加模型的多样性。每个决策树独立地对数据学习,最终通过投票(分类问题)或者平均(回归问题)的方式整合各个决策树的预测结果。随机森林具有较强的鲁棒性,能够处理复杂

的非线性关系,同时对于噪声和过拟合具有较好的抵抗性。

BP神经网络是一种常见的人工神经网络,由输入层、隐藏层和输出层组成,每层包含多个神经元,如图1所示。训练过程通过反向传播算法,即误差反向传递,不断调整网络权重,以期使网络的实际输出值和期望输出值的均方误差最小。BP神经网络能够学习复杂的非线性映射关系,适应性较强,被广泛用于模式识别、预测和控制等领域。

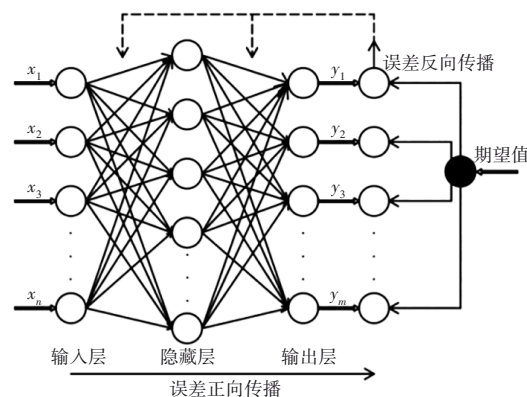


图1 BP神经网络结构示意图

Fig.1 BP neural network structure diagram

LSTM神经网络是循环神经网络的改良型,旨在解决传统循环神经网络在长序列训练过程中的梯度消失和梯度爆炸等问题。典型的LSTM网络结构见图2,由遗忘门、输入门、更新记忆、输出门组成,控制着信息的取舍、输入、更新、输出,能够更好地捕捉和处理时间序列中的长期依赖关系。

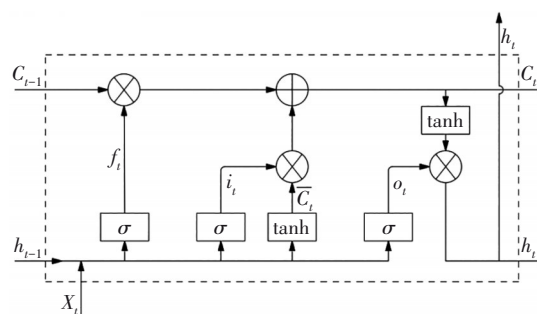


图2 LSTM网络结构示意图

Fig.2 LSTM network structure diagram

1.2 数据及预处理

本研究采用W市某区4个污水检查井布设的在线液位计(WS1~WS4)监测的实时液位数据,数据采集间隔为1 min,采集时间为2023年1月1日—12月22日,共356 d。考虑到污水检查井每分钟液位变化幅度较小,为更高效地训练时间序列预测模型,将每小时监测液位数据的最大值作为一个样本数据,记为每小时液位峰值 $L_{\max}$,数据间隔为1 h,共8 544个液位峰值数据。以WS1在线液位计为例,其监测的液位峰值数据如图1所示。

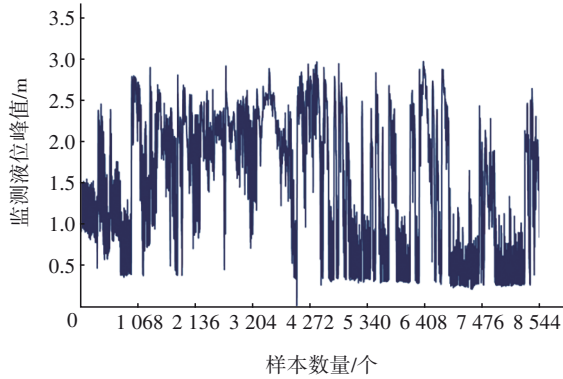


图3 WS1在线液位计监测的液位峰值数据

Fig.3 Peak liquid level data monitored by WS1 on-line liquid level gauge

在进行液位峰值预测模型训练时,首先将时间序列数据进行预处理,通过预处理将监测的液位峰值数据形成训练所需要的数据集,将液位峰值预测问题转化为一个包含输入序列和输出序列的监督学习问题,使用监督学习算法对历史监测值进行训练以预测未来的观测值。为了把问题从无监督转化为有监督,将每24个相邻时间步 $t+0, t+1, \dots, t+22, t+23$ 的液位峰值划分为一个样本作为模型输入,来预测第 $t+24$ 步的液位峰值,即将过去24 h的液位峰值作为模型的输入预测第25小时的液位峰值作为模型的输出。

对监测的液位峰值数据集进行标准化处理后,将60%的数据集作为训练集,用于训练模型的参数和权重,让模型能够逐步学习数据的特征和模式;将20%的数据集作为验证集,用于调整模型的超参数,选择最佳模型;将剩余20%的数据集作为测试集,用于评估训练好的模型在未知数据上的预测能力和泛化能力。

1.3 模型评价指标

为评估预测模型在不同方面的性能表现,选择

MAE、RMSE和 R^2 作为评价指标。MAE和RMSE从两个不同的角度反映预测误差的大小,全面了解模型对异常值的适应能力; R^2 反映预测模型在整体拟合方面的表现, R^2 值越接近1说明模型的拟合效果越优。3个评价指标的计算公式如下:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |L_i - \hat{L}_i| \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (L_i - \hat{L}_i)^2} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (L_i - \hat{L}_i)^2}{\sum_{i=1}^n (L_i - \bar{L}_i)^2} \quad (3)$$

式中: n 为样本数量; L_i 为液位峰值的实测值; $\bar{L}_i$ 为液位峰值的实测均值; $\hat{L}_i$ 为液位峰值的预测值。

2 超参数优选

由于不同数据集具有不同的数据特性,如数据分布、数据频率等,这些特性会影响预测模型的性能表现,为了全面评估模型的泛化能力并挑选出最优模型,本研究使用网格搜索法作为超参数优选方法,精确地识别出表现最佳的模型配置,以确保充分评估每个数据集的不同模型性能。

2.1 RF模型超参数优选

RF模型超参数优选主要通过调整决策树数 N_{RF} 、决策树最大深度 D_{RF} 、最小叶子节点样本数 M_{RF} 三个超参数在给定区间内进行取值来组成多个超参数组合,对每个超参数的组合进行模型训练,并采用验证集的模型评价指标来选择最佳的超参数配置。具体来说, D_{RF} 取值区间为[20, 50, 100, 150, 200], N_{RF} 取值区间为[50, 100, 150, 200, 300, 400, 500], M_{RF} 取值区间为[5, 10, 15, 20]。

以WS1污水检查井为例,通过网格搜索法训练140种超参数组合的RF模型,并采用 R^2 指标评估模型性能,以此来选择最优超参数组合。将RF模型超参数优选结果用三维散点图展示,并将使 R^2 最大的超参数配置突出显示,可通过颜色深浅来判断模型在不同超参数组合下的性能表现,结果见图4。可以看出,WS1的RF模型的最佳 $N_{\text{RF}}=100$ 、 $D_{\text{RF}}=200$ 、 $M_{\text{RF}}=10$,而当 N_{RF} 超过200或 $N_{\text{RF}}=200$ 且 $D_{\text{RF}}=200$ 时,拟合较差。同理可得其他污水检查井液位峰值数据的RF模型最优超参数组合如下:WS2, $N_{\text{RF}}=100$ 、 $D_{\text{RF}}=200$ 、 $M_{\text{RF}}=5$; WS3, $N_{\text{RF}}=200$ 、 $D_{\text{RF}}=200$ 、 $M_{\text{RF}}=5$;

WS4, $N_{RF}=100$ 、 $D_{RF}=200$ 、 $M_{RF}=5$ 。

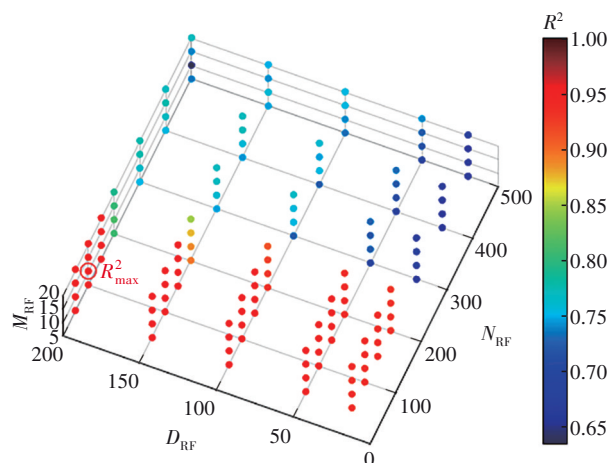


图4 WS1的RF模型超参数优选三维散点图

Fig.4 3D scatter plots of RF model hyper parameter optimization for WS1

2.2 BP神经网络模型超参数优选

BP神经网络模型的激活函数采用Sigmoid函数,训练轮数为100,采用随机权重初始化方法,通过在给定区间内调整隐藏层神经元数 H_{BP} 、学习率 L_{BP} 两个超参数来训练模型和选择最佳超参数配置,具体来说, L_{BP} 取值区间为 $[0.001, 0.01, 0.1]$, H_{BP} 取值区间为 $[50, 100, 150, 200, 250, 300]$ 。

以WS2污水检查井为例来进行BP神经网络模型的超参数优选,以MAE为评价指标来衡量模型性能,超参数优选结果见图5。可以看出,最优超参数组合为 $H_{BP}=50$ 、 $L_{BP}=0.1$ 。同理可得其他污水检查井的BP神经网络模型最优超参数组合如下:WS1, $H_{BP}=300$ 、 $L_{BP}=0.001$; WS3, $H_{BP}=50$ 、 $L_{BP}=0.1$; WS4, $H_{BP}=300$ 、 $L_{BP}=0.001$ 。

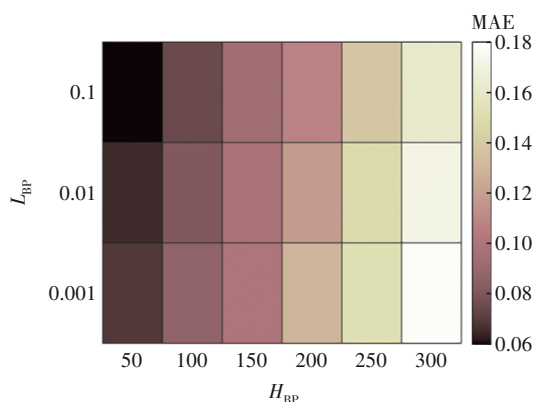


图5 WS2的BP神经网络模型超参数优选结果

Fig.5 Hyper parameter optimization results of BP neural network model for WS2

2.3 LSTM神经网络模型超参数优选

LSTM神经网络在时间序列预测问题中应用广泛,本研究基于LSTM神经网络建立对污水检查井液位峰值的预测模型。首先定义网络的层结构,包括输入层、LSTM层、dropout层、ReLU激活层、全连接层和回归层,为防止过拟合,dropout取0.1。为确定LSTM网络的最优超参数,训练由不同LSTM隐藏层单元数量 H_{LSTM} 、不同学习率 L_{LSTM} 、不同最大训练轮数 T_{LSTM} 三个超参数相互组合成60个LSTM模型。具体来说, H_{LSTM} 取值区间为 $[50, 100, 150, 200]$, L_{LSTM} 取值区间为 $[0.1, 0.01, 0.001]$, T_{LSTM} 取值区间为 $[10, 20, 30, 40, 50]$ 。经过优选,可得到LSTM神经网络模型的最优超参数组合如下:WS1, $H_{LSTM}=100$ 、 $L_{LSTM}=0.1$ 、 $T_{LSTM}=40$; WS2, $H_{LSTM}=150$ 、 $L_{LSTM}=0.01$ 、 $T_{LSTM}=50$; WS3, $H_{LSTM}=200$ 、 $L_{LSTM}=0.01$ 、 $T_{LSTM}=50$; WS4, $H_{LSTM}=150$ 、 $L_{LSTM}=0.1$ 、 $T_{LSTM}=30$ 。

3 结果与讨论

基于模型超参数优选结果,依据4个污水检查井液位峰值监测数据分别建立基于RF、BP神经网络、LSTM神经网络的3种最优预测模型,对比不同模型在测试集的性能表现。考虑到机器学习和深度学习每次运行会产生不同的预测结果,为尽可能合理评估不同模型的性能,对每种预测模型都运行5次并对测试集各个评价指标取平均值,不同预测模型的性能对比如表1所示。以WS1污水检查井为例,液位峰值实测值与3种模型的预测值对比如图6所示。从表1和图6均可以看出,3种模型在预测液位峰值时拟合程度较好,预测效果较为稳定,能够较好地解释数据的变化趋势。

以污水检查井WS1为例,3种预测模型的 R^2 值分别为0.946、0.943、0.942,在整体拟合程度方面3种预测模型虽然存在细微的差异,但均具有较好的拟合程度,都能较好地捕捉到数据中的主要趋势和特性;在预测误差和稳定性表现方面,以污水检查井WS3为例,由表1可知, $MAE_{LSTM} < MAE_{BP} < MAE_{RF}$ 、 $RMSE_{LSTM} < RMSE_{BP} < RMSE_{RF}$,相对于RF模型,BP和LSTM神经网络模型具有更小的MAE和RMSE,这意味着BP和LSTM神经网络模型对异常值的适应能力更强,表明这两种模型在面对不同情况时具有更优的泛化能力,其中LSTM神经网络模型的预测误差最小、预测结果最为稳定。

表 1 不同预测模型的性能对比

Tab.1 Performance comparison of different prediction models

项目	RF 模型			BP 神经网络模型			LSTM 神经网络模型		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
WS1	0.119	0.189	0.946	0.119	0.192	0.943	0.044	0.065	0.942
WS2	0.069	0.129	0.985	0.060	0.127	0.986	0.018	0.029	0.986
WS3	0.069	0.121	0.986	0.047	0.093	0.992	0.019	0.029	0.987
WS4	0.052	0.108	0.983	0.042	0.080	0.990	0.019	0.035	0.980

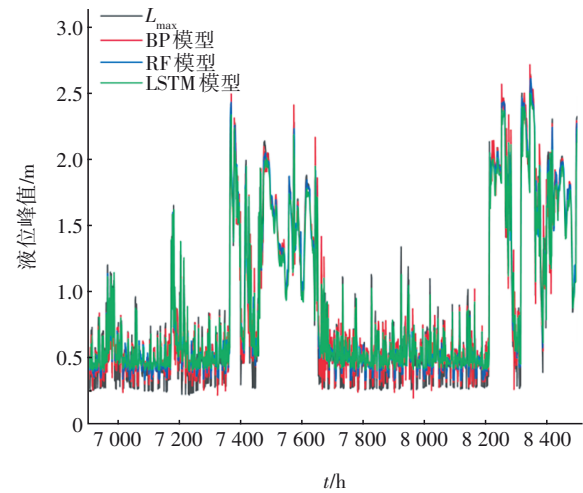


图 6 WS1 的液位峰值监测值与预测值对比

Fig.6 Comparison of monitored and predicted values of peak liquid level for WS1

为了清晰展示不同模型对同一数据的预测结果差异,以 WS4 污水检查井 2023 年 7 月 1 日—7 日的液位峰值监测数据为例,3 种模型的预测结果见图 7。

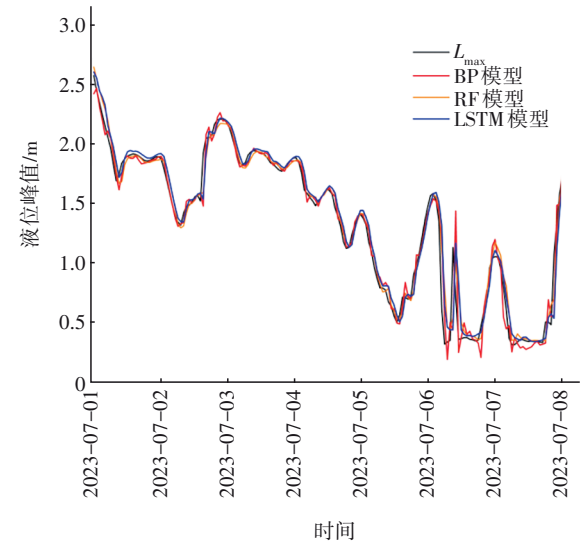


图 7 不同模型对 WS4 液位峰值的预测结果对比

Fig.7 Comparison of prediction results of peak liquid level for WS4 using different models

从图 7 可以看出,3 种模型对 WS4 污水检查井的液位峰值预测结果与实际监测值 $L_{\max}$ 的吻合程度较高,能够精准捕捉到时间序列数据的趋势和变化,表现出可靠的预测效果;但在峰值液位较小时,与 LSTM 神经网络模型相比,RF 模型和 BP 神经网络模型的预测值与实际监测值之间存在相对较大的误差与不稳定性。

由表 1 可知,3 种模型在对 WS1~WS4 的液位峰值分别进行预测时,WS1 的 MAE 和 RMSE 最大,这说明 3 种模型在对 WS1 的液位峰值进行预测时误差较大且不稳定。同时,当液位峰值连续处于较低点时,WS1 的液位峰值预测值与监测值曲线波动差异较大(见图 6),而 WS2~WS4 的液位峰值预测值曲线与监测值曲线呈现较为一致的变化趋势,因此不同数据集由于数据特性不同可能对预测结果有一定的影响。

为了探究不同数据集对同一模型预测结果的影响程度,以 LSTM 神经网络模型为例,不同数据集下该模型模拟结果的评价指标值差异如图 8 所示。

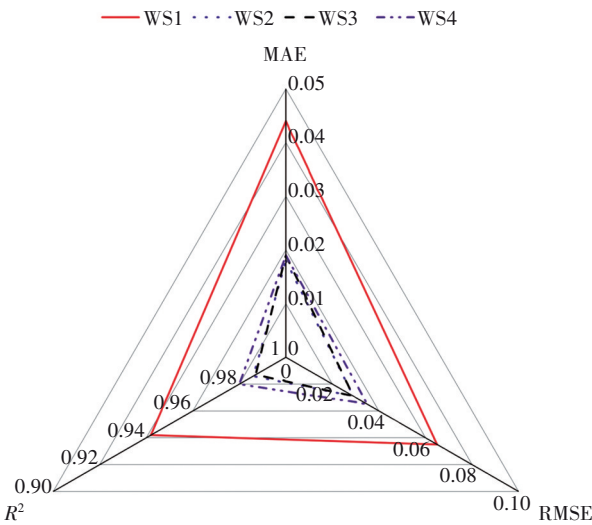


图 8 不同数据集对 LSTM 神经网络模型预测结果的影响

Fig.8 Influence of different datasets on prediction results of LSTM neural network models

雷达图面积 S 越小表示基于该数据集所建立的LSTM神经网络模型性能表现越好。从图8可以看出, $S_{WS1} \gg S_{WS4} > S_{WS3} > S_{WS2}$,WS1的雷达图面积最大且远大于其他数据集,WS2~WS4的雷达图面积接近、仅存在细微差异。因此,不同数据集对预测结果存在一定程度的影响,其中基于WS1数据集所建立的LSTM神经网络预测模型性能表现相对较差,WS2~WS4的LSTM神经网络预测模型性能表现则相对较好。

4 结论

基于W市某区4个污水检查井液位计2023年1月1日—12月22日的在线监测数据,通过对数据预处理后以网格搜索法作为超参数优选方法,分别构建基于RF、BP神经网络、LSTM神经网络3种方法的液位峰值最优预测模型,并以MAE、RMSE、 R^2 作为评价指标对模型预测性能进行比较分析。结果表明,在预测污水检查井液位峰值时,不同数据集对预测结果存在一定程度的影响;基于同一数据集所建立的3种预测模型均表现出较好的拟合程度,但在误差与稳定性表现方面有所差异,其中LSTM神经网络模型的表现优于RF和BP神经网络模型。LSTM神经网络模型的预测结果稳定,具有较优的泛化能力,是一种有效的污水检查井液位峰值预测方法。

参考文献:

- [1] 石小峰,马宏伟,段琦琦,等. 皖北某城市排水管网调查及整改对策分析[J]. 工业用水与废水, 2018, 49(4): 36-39.
SHI Xiaofeng, MA Hongwei, DUAN Qiqi, *et al.* Analysis on investigation of urban drainage pipe network in a city of northern Anhui Province and its rectification measures [J]. Industrial Water & Wastewater, 2018, 49(4): 36-39 (in Chinese).
- [2] 曹业始, KROISS H, VAN LOOSDRECHT M, 等. 进一步改善城市网-厂-水体污水系统升级改造和运行效率的四点看法[J]. 给水排水, 2024, 50(4): 137-144.
CAO Yesi, KROISS H, VAN LOOSDRECHT M, *et al.* The four issues relating to improve the efficiency of upgrading and operation of urban sewage systems in China [J]. Water & Wastewater Engineering, 2024, 50(4): 137-144 (in Chinese).
- [3] 张嘉治,张雷,郑丙辉,等. 基于频次分析法的污水处理厂在线监测预警技术研究[J]. 中国环境科学, 2012, 32(10): 1792-1798.
ZHANG Jiazhi, ZHANG Lei, ZHENG Binghui, *et al.* Monitoring and early alarming technology of sewage treatment plant based on frequency analysis [J]. China Environmental Science, 2012, 32(10): 1792-1798 (in Chinese).
- [4] REN T, LIU X F, NIU J W, *et al.* Real-time water level prediction of cascaded channels based on multilayer perception and recurrent neural network [J]. Journal of Hydrology, 2020, 585: 124783.
- [5] ZHANG D, HOLLAND E S, LINDHOLM G, *et al.* Hydraulic modeling and deep learning based flow forecasting for optimizing inter catchment wastewater transfer[J]. Journal of Hydrology, 2018, 567: 792-802.
- [6] 何媛滨,方正,康丹,等. 基于BP神经网络的合流制截污管溢流实时控制方法[J]. 中国给水排水, 2024, 40(3): 120-123.
HE Yuanbin, FANG Zheng, KANG Dan, *et al.* Real-time control method for combined sewer overflow based on BP neural network [J]. China Water & Wastewater, 2024, 40(3): 120-123 (in Chinese).
- [7] 李双宇,张明凯,刘艳臣,等. 基于LSTM模型的排水系统流量预测研究[J]. 中国给水排水, 2022, 38(5): 59-64.
LI Shuangyu, ZHANG Mingkai, LIU Yanchen, *et al.* Flow prediction of drainage system based on long short time memory model [J]. China Water & Wastewater, 2022, 38(5): 59-64 (in Chinese).

作者简介:杨利刚(2000—),男,安徽阜阳人,硕士研究生,主要研究方向为排水管网外水评估及算法。

E-mail:25085351@qq.com

收稿日期:2024-05-29

修回日期:2024-07-28

(编辑:刘贵春)