DOI:10. 19853/j. zgjsps. 1000-4602. 2022. 05. 009

基于LSTM模型的排水系统流量预测研究

李双宇[1,2]， 张明凯[2,3]， 刘艳臣[3]， 施汉昌[2,3]
（1. 北京大学 工学院，北京 100871；2. 北京协同创新研究院，北京 100094；3. 清华大学 环境学院，北京 100084）

摘 要： 排水系统流量预测对于城市水安全、污水厂优化运行具有重要意义。与需要复杂建模和大量地理信息数据的传统水文水力学模型不同，机器学习可以通过数据驱动实现排水系统的流量预测预警。结合流量数据的时序性，分别在单变量（流量）、双变量（流量和降雨）的情况下，采用5种长短期记忆神经网络（LSTM）模型（Vanilla LSTM、Stacked LSTM、Bidirectional LSTM、CNN LSTM、ConV LSTM）对江苏省无锡市某污水处理厂的进水流量进行预测。结果表明，Bidirectional LSTM最优的实验参数条件是：隐藏层单元数为250，训练轮数为200，训练集样本数为250；在同等条件下，Bidirectional LSTM相较其他4种方法可以更有效地预测未来流量；相比仅输入流量变量，在增加降雨变量后，可以提升近20%的流量预测精度。

关键词： 排水系统； LSTM模型； 流量预测； 时间序列； 最优实验参数
中图分类号： TU992 **文献标识码：** A **文章编号：** 1000－4602（2022）05－0059－06

Flow Prediction of Drainage System Based on Long Short Time Memory Model

LI Shuang-yu[1,2], ZHANG Ming-kai[2,3], LIU Yan-chen[3], SHI Han-chang[2,3]
(1. *College of Engineering*, *Peking University*, *Beijing* 100871, *China*; 2. *Beijing Institute of Collaborative Innovation*, *Beijing* 100094, *China*; 3. *School of Environment*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: Flow prediction of drainage systems is of great significance for urban water safety and optimal operation of wastewater treatment plants. Different from traditional hydrological models which need complex modeling and a large amount of geographic information data, machine learning can realize flow prediction and early warning of a drainage system through data driving. In combination with the time sequence of flow data, five long short time memory (LSTM) models (Vanilla LSTM, Stacked LSTM, Bidirectional LSTM, CNN LSTM and ConV LSTM) under the conditions of single variable (flow) and double variable (flow and rainfall) were applied to predict the inlet flow of a wastewater treatment plant in Wuxi City, Jiangsu Province. In the parameter selection experiment, the optimal parameter condition of Bidirectional LSTM was that the number of LSTM hidden layer units, training epochs and training set samples were 250, 200 and 250. Under the same condition, Bidirectional LSTM predicted the future flow more effectively than the other four methods. Compared with simulation with flow as the only variable, its accuracy of flow prediction was improved by nearly 20% after adding rainfall as another variable.

Key words: drainage system; LSTM model; flow prediction; time sequence; optimal

基金项目： 国家水体污染控制与治理科技重大专项（2017ZX07103007）
通信作者： 张明凯 E-mail：zhangmk@bici.org

experimental parameter

排水系统(污水系统、雨水系统、混流制系统)的流量预测对于城市内涝预警、溢流污染控制、泵站调度都有重要意义。随着我国城市化进程的快速发展,排水系统的管道铺设长度逐年递增,雨水、污水收集率都在不断提高。但许多城市仍存在排水设施不完善、管网配套建设滞后、排放能力不够、管网老化严重、内涝及溢流等问题,导致了严重的环境污染,并阻碍了城市发展。有效的流量预测模型,可以为排水系统的优化运行提供保障,减少溢流、内涝等事件发生。

传统的流量预测一般采用SWMM、MIKE URBAN、InfoWorks ICM 等水文水力学模型及软件。但这些方法都需要详细的地理信息数据(汇水区下垫面属性、土地利用类型、土壤渗透系数、管网坡度、管径、粗糙度等)和水文气象数据(降雨、融雪、潮汐等)、复杂的建模过程,以及大量的参数率定和校准工作。

人工智能作为新兴交叉学科,近年来得到了飞速发展和应用。许多神经网络方法如RNN、LSTM、CNN、GAN 等,在排水系统流量预测方面实现了诸多重要应用[1]。Ren 等[2]采用多层感知器模型和RNN 模型,对多级串联管道的液位变化进行预测。Zhang 等[3]研究发现,LSTM 模型能够有效用于管网流量预测;在其对溢流监测的研究中,LSTM 和GRU 在多步超前时间序列预测方面具有优越性能[4];在对挪威拉门市污水在线存储控制系统管理的研究中,比较了 Elman、NARX 和LSTM 三种神经网络,实验结果表明LSTM 具有良好的时间序列预测能力[5]。Karimi 等[6]于2019 年分别采用人工神经网络 ANN、长短期记忆网络LSTM、LASSO 3种机器学习方法对排水系统进行了流量预测,结果表明在同等实验设置条件下,LSTM 的流量预测效果最好。

以上研究分析了 LSTM、NARX、GRU 等不同类型的神经网络进行流量预测的效果,得出 LSTM 模型在时序数据预测方面具有更好的性能。但对如何提升 LSTM 的预测性能和对 LSTM 自身及其相关改进缺乏更深入的研究。笔者使用5种不同的LSTM 模型(Vanilla LSTM、Stacked LSTM、ConV LSTM、CNN LSTM、Bidirectional LSTM),以江苏省无锡市某污水处理厂实际进水数据为例,模拟预测排水系统的流量变化,旨在为排水系统的实际运行提供参考。

1　材料和方法

1.1　LSTM 模型介绍

相比经典 LSTM,Vanilla LSTM 有一个全连接隐藏层和一个用于进行预测的全连接输出层,在添加窥视孔连接(控制门前的存储单元信息)后,使门不仅依赖于先前的隐藏状态 h_t,而且还依赖于先前的内部状态 C_t。

经典 LSTM 模型由一个隐藏的 LSTM 层和一个标准的前馈输出层组成,Stacked LSTM 在其基础上进行了扩展,将多个隐藏的 LSTM 层堆叠在另一层上,其中每个层包含多个内存单元。Stacked LSTM 使用堆叠层来创建输入数据的分层特征,这种堆叠的隐藏层可以增加模型的复杂性,从而让计算更加准确。

Bidirectional LSTM 可以很好地解决流量数据与前后时刻均相关的问题。图 1 为 Bidirectional LSTM 网络结构,当正向计算时,前馈层从 1 时刻到 t 时刻正向计算一遍,得到并保存每个时刻向前隐藏层的输出;当反向计算时,反馈层沿着 t 时刻到 1 时刻反向计算一遍,得到并保存每个时刻向后隐藏层的输出,在每个时刻结合前馈层和反馈层的相应时刻输出结果,得到最终的输出 o_t。

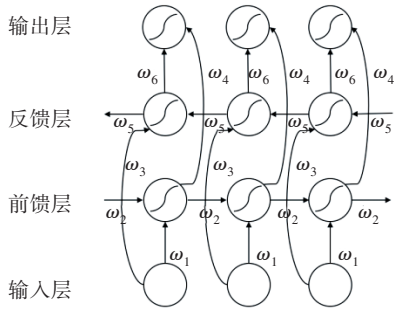


图 1　Bidirectional LSTM 网络结构

Fig.1　Network structure of Bidirectional LSTM

CNN 模型可用于 LSTM 为后端的混合模型中,这种混合模型被称为 CNN LSTM,它可以被看成两个子模型:在输入数据中使用 CNN 层进行特征提取,之后结合 LSTM 来进行序列预测。

ConV LSTM 和 CNN LSTM 的主要区别在于前者

仅对于输入进行卷积计算，其中输入的卷积读数直接嵌入到每个LSTM单元中，以此来捕获基础空间特征。

1.2　数据及预处理

本研究的数据为江苏省无锡市某污水处理厂的进水流量，采集时间间隔为1 h，共552条，如图2所示。其中进水流量均值为8 790.75 m^3/h，最小值为7 705.44 m^3/h，最大值为12 416.09 m^3/h；降雨量最大值为18.5 mm/h。在无降雨情况下，污水厂的进水流量呈现明显的旱季周期变化规律，当降雨事件发生后，污水厂进水流量显著升高，随后逐渐降低。

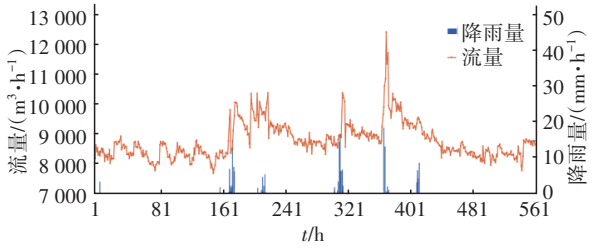


图2　进水流量与区域降雨量的变化

Fig.2　Change of influent flow and regional rainfall

时间序列数据在训练LSTM神经网络前需要进行预处理，将其转换成具有输入和输出分量的样本，使问题从无监督转化为有监督。本研究将每4个相邻时间步t+0、t+1、t+2、t+3的降雨和流量划分为一个样本作为模型输入，从而预测第t+4步的流量。运算时，CNN LSTM和ConV LSTM将每个样本又分成两个子样本，每个子样本为2个相邻时间步的降雨和流量数值。

2　结果与讨论

2.1　评价指标

由于MSE、RMSE、MAE在衡量上不具有上限，为了更好地量化预测方法的性能，本研究使用了R^2，能够将模型的预测精度与普通基准模型的精度进行比较，见式(1)。

$$R^2 = 1 - \frac{\sum_{t=1}^{n}\left(\hat{y}_t - y_t\right)^2}{\sum_{t=1}^{n}\left(\bar{y} - y_t\right)^2} = 1 - \frac{\left[\sum_{t=1}^{n}\left(\hat{y}_t - y_t\right)^2\right]/n}{\left[\sum_{t=1}^{n}\left(\bar{y} - y_t\right)^2\right]/n} = 1 - \frac{\mathrm{MSE}(\hat{y},y)}{\mathrm{Var}(y)} \tag{1}$$

式中：y_t为实际值；$\hat{y}_t$为预测值；$\bar{y}$为y_t值的平均值；n为数据样本的总数；R^2取值为$(-\infty,1]$，如果R^2=1，表示是一个非常好的预测，$R^2=-\infty$则相反。R^2越大表明预测精度越高，如果预测方法R^2>0.5，则该方法是成功的。

2.2　单变量(流量)预测结果分析

考虑到算法数值精度的差异，结果会略有不同，为消除实验误差，每组实验运行5次，并比较平均结果与所获结果的方差。

实验一探究了隐藏层中LSTM单元数对结果的影响，按照间隔50在隐藏层中分别设置了100~350个LSTM单元，依次进行6组实验，表1展示了不同方法下的R^2及方差。

表1　隐藏层中LSTM单元数对实验精度的影响

Tab.1　Influence of the number of LSTM hidden layer units on experimental accuracy

项　目		单元数					
LSTM方法	评价指标	100	150	200	250	300	350
Vanilla	R^2	0.691 3	0.683 6	0.694 0	0.687 5	0.688 4	0.663 8
	VAR	0.000 3	0.000 2	0.000 5	0.000 6	0.000 1	0.000 8
Stacked	R^2	0.634 4	0.641 9	0.645 7	0.649 6	0.663 6	0.621 1
	VAR	0.002 3	0.000 7	0.000 6	0.000 1	0.000 1	0.001 7
Bidirectional	R^2	0.754 7	0.741 7	0.737 0	0.752 5	0.726 6	0.725 4
	VAR	0.000 6	0.000 5	0.000 5	0.000 7	0.000 9	0.000 5
CNN	R^2	0.715 4	0.711 7	0.702 1	0.700 2	0.689 6	0.697 3
	VAR	0.000 2	0.000 1	0.000 0	0.000 1	0.001 1	0.000 0
ConV	R^2	0.718 0	0.715 1	0.720 5	0.723 4	0.714 0	0.726 7
	VAR	0.000 1	0.000 6	0.000 4	0.000 5	0.000 1	0.000 3

对实验结果进行分析发现，单元数为100~300时，随着LSTM单元数的增多，Vanilla LSTM预测精度均在0.68~0.70之间，相对比较稳定；当训练单元数为350时，预测精度显著降低。Stacked LSTM预测精度随着单元数的增加先升高后下降，且当单元数为300时预测精度最佳。在单元数为100~250时，Bidirectional LSTM预测精度稳定在0.73~0.76范围内；当单元数为300、350时，精度显著降低。分析原因，在一定范围内，单元数增多能够提高网络训练效率，降低误差，但单元数持续增加会使网络节点复杂化，产生过拟合，从而降低模型的预测精度。对于同种方法，最优结果已用方框标出，当LSTM单元数为100时，Bidirectional LSTM、CNN LSTM达到最佳预测精度，分别为0.754 7、0.715 4；

当单元数为200时，Vanilla LSTM达到最佳预测精度，为0.694 0；当单元数为300时，Stacked LSTM达到最佳预测精度，为0.663 6；当单元数为350时，ConV LSTM达到最佳预测精度，为0.726 7。

在隐藏层数相同的情况下，Bidirectional LSTM的预测精度明显高于其他方法。其优秀的双向计算能力，可以同时完成正反向计算，在面对突发降雨情况时，能够融合降雨与非降雨事件，从而有效完成预测。

实验二从100到300按照间距为50设置5组实验，以探究训练轮数对预测精度的影响。通过表2可以看到，随着训练轮数的增加，Vanilla LSTM、Bidirectional LSTM方法的精度在训练轮数为200时，达到最佳预测精度，分别为0.805 7、0.804 6；Stacked LSTM、ConV LSTM在训练轮数为100时达到了最优预测精度，分别为0.799 9和0.799 4；当训练轮数为300时，CNN LSTM模型的预测精度最佳，为0.757 9。当训练轮数固定时进行不同实验方法的比较，可见Bidirectional LSTM在5组实验中的3组都取得了最佳精度。从实验结果还可以发现，相比其他4种方法，Stacked LSTM模型的方法方差较大，但总体影响不大。

表2　训练轮数对实验精度的影响

Tab.2　Influence of the number of training epochs on experimental accuracy

项　目		训练轮数				
LSTM方法	评价指标	100	150	200	250	300
Vanilla	R^2	0.792 9	0.716 6	0.805 7	0.768 0	0.792 9
	VAR	0.000 5	0.000 1	0.000 0	0.000 1	0.000 2
Stacked	R^2	0.799 9	0.698 3	0.788 8	0.678 9	0.796 6
	VAR	0.000 1	0.001 0	0.000 2	0.010 3	0.000 3
Bidirectional	R^2	0.801 9	0.771 5	0.804 6	0.773 4	0.765 9
	VAR	0.000 0	0.000 0	0.000 0	0.000 2	0.003 4
CNN	R^2	0.741 7	0.711 0	0.747 0	0.735 0	0.757 9
	VAR	0.001 0	0.000 8	0.000 2	0.000 2	0.000 1
ConV	R^2	0.799 4	0.724 8	0.798 7	0.753 6	0.772 7
	VAR	0.000 1	0.000 1	0.000 0	0.000 7	0.001 0

实验三探究了训练集样本数对预测精度的影响，结果见表3。可知，随着训练集样本数的增多，相较于样本数较少的实验，Vanilla LSTM在样本数为250、300时，预测精度较高，且样本数为250时预测精度最大，为0.746 8，而当样本数为350时，预测精度明显下降；Stacked LSTM在使用100~200个样本预测时，精度基本保持稳定，在使用250个样本预测时精度显著提升，且在使用300个样本预测时达到最佳，为0.744 1，之后保持相对稳定；ConV LSTM预测精度随样本数的增加先上升后下降，在训练集样本数为300时，精度达到最佳，为0.754 3；CNN LSTM随样本数的增加，预测精度先升后降，当样本数为250时预测精度最大，为0.751 3；Bidirectional LSTM受样本数的影响不大，基本可以保持训练精度的稳定性，样本数为250时预测精度最大，为0.779 3。这是因为样本较少时，所获取的信息太少，随着样本的增加，可以提供更多信息，为模型训练提供助力，但样本达到一定数量时，增加样本会导致模型过拟合，训练时间增加，导致模型泛化能力差。

表3　训练集样本数对实验精度的影响

Tab.3　Influence of the number of training set samples on experimenal accuracy

项　目		训练集样本数					
LSTM方法	评价指标	100	150	200	250	300	350
Vanilla	R^2	0.676 5	0.659 4	0.701 6	0.746 8	0.745 8	0.727 2
	VAR	0.000 2	0.003 8	0.000 3	0.001 0	0.000 3	0.001 1
Stacked	R^2	0.662 9	0.669 4	0.662 5	0.711 0	0.744 1	0.743 2
	VAR	0.000 0	0.002 2	0.000 1	0.001 2	0.029 0	0.000 7
Bidirectional	R^2	0.749 3	0.777 8	0.727 0	0.779 3	0.761 0	0.762 1
	VAR	0.000 2	0.000 1	0.000 8	0.000 4	0.000 7	0.000 5
CNN	R^2	0.694 5	0.711 9	0.716 3	0.751 3	0.734 6	0.745 5
	VAR	0.000 1	0.001 9	0.000 1	0.000 1	0.000 1	0.000 0
ConV	R^2	0.617 4	0.673 9	0.708 6	0.732 4	0.754 3	0.719 2
	VAR	0.017 4	0.000 4	0.001 2	0.000 3	0.000 0	0.001 2

综上可知，Bidirectional LSTM模型在全局中预测精度较高，对于极值点的拟合能力较好，且泛化能力强，能够有效预测降雨与无降雨条件下的流量情况。为获取最佳预测，设置LSTM的隐藏层单元数为250、训练轮数为200、训练集样本数为250。图3为使用Bidirectional LSTM得到的5次预测均值与真实测量值。

从图3可以发现以下两点：第一，Bidirectional LSTM在预测过程中存在明显的滞后性；第二，算法的整体预测处在一个平稳状态，但当进水流量增大以后，波峰预测存在一定误差。以上两点均与实验的性质有关，实验使用前4个样本预测第5个样本，

所以预测结果会高度依赖前4个输入样本，从而对实验精度产生滞后影响，波峰为区间内的最大值，通过较小的样本预测较大的样本也较难获得准确数值。

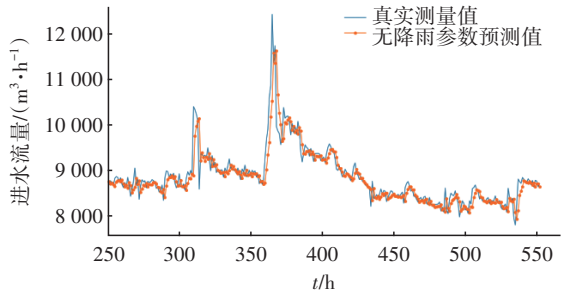


图3　进水流量预测结果与真实测量值

Fig.3　Prediction results and real values of influent flow

2.3　双变量（降雨、进水流量）预测结果分析

在增加降雨的情况下，选定双变量下5种LSTM的相关参数（设定隐藏层单元数为250、训练轮数为300、训练样本数为200），对进水流量进行5次预测实验，结果见表4。

表4　5次实验结果

Tab.4　Results of five experiments

项　目	Vanilla	Stacked	Bidirectional	CNN	ConV
1	0.883 0	0.803 3	0.915 2	0.859 6	0.860 1
2	0.836 4	0.842 1	0.973 9	0.899 5	0.948 3
3	0.864 5	0.802 1	0.988 6	0.883 9	0.849 6
4	0.911 8	0.901 9	0.951 6	0.908 9	0.827 8
5	0.849 4	0.870 4	0.989 7	0.919 2	0.886 3
平均值	0.869 0	0.844 0	0.963 8	0.894 2	0.874 4
方差	0.000 9	0.001 9	0.001 0	0.000 5	0.002 1

从表4可以看出，Bidirectional LSTM相比其他方法有更高的预测精度，5次平均预测结果为0.963 8，CNN LSTM和ConV LSTM的预测精度也较高，5次平均预测结果分别为0.894 2和0.874 4。

为了清晰地展示预测结果，表5给出了不同方法下第451~461个样本点的5次实验平均预测结果。通过计算预测值与真实测量值之间差值的绝对值，来判定预测是否接近真实测量值。从单个样本预测最接近真实测量值的个数来看，Bidirectional LSTM方法在11个样本预测中有4个，Stacked LSTM有3个，ConV LSTM有2个，Vanilla LSTM和CNN LSTM均有1个。因此，无论从总体预测还是单一样本点预测，Bidirectional LSTM的预测效果均优于其他4种LSTM方法。

表5　不同方法在第451~461个样本点的预测结果

Tab.5　Prediction results of different methods at 451–461 sample points　　m^3/h

样本点	测量值	Vanilla	Stacked	Bidirectional	CNN	ConV
451	8 437.5	8 509.7	8 461.0	8 418.8	8 457.9	**8 424.4**
452	8 478.0	8 512.0	8 494.7	8 452.5	**8 467.4**	8 437.1
453	8 524.3	8 552.2	**8 542.5**	8 503.0	8 498.3	8 473.8
454	8 284.1	8 461.9	8 321.5	**8 273.3**	8 499.3	8 357.6
455	8 625.6	8 512.7	**8 616.1**	8 580.2	8 582.1	8 476.6
456	8 718.2	8 706.7	8 740.3	**8 708.3**	8 643.8	8 614.8
457	8 755.8	**8 767.9**	8 778.3	8 722.3	8 702.5	8 674.3
458	8 663.2	8 761.9	8 692.6	8 643.3	8 733.2	**8 666.8**
459	8 547.5	8 673.2	8 575.0	**8 528.7**	8 684.3	8 586.3
460	8 590.9	8 637.9	**8 605.1**	8 568.5	8 635.7	8 569.3
461	8 437.5	8 576.9	8 468.3	**8 428.0**	8 605.2	8 482.2

图4为第420~500个样本点的预测情况，方法预测的总体拟合效果均可以接受。值得注意的是，Bidirectional LSTM方法相比其他方法在预测过程中的预测效果最好。

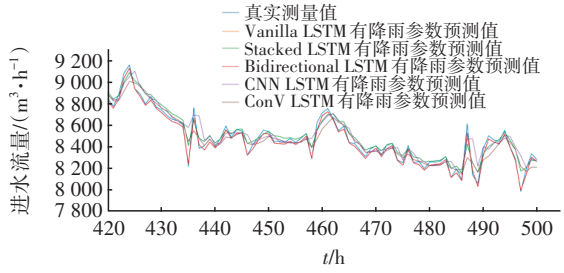
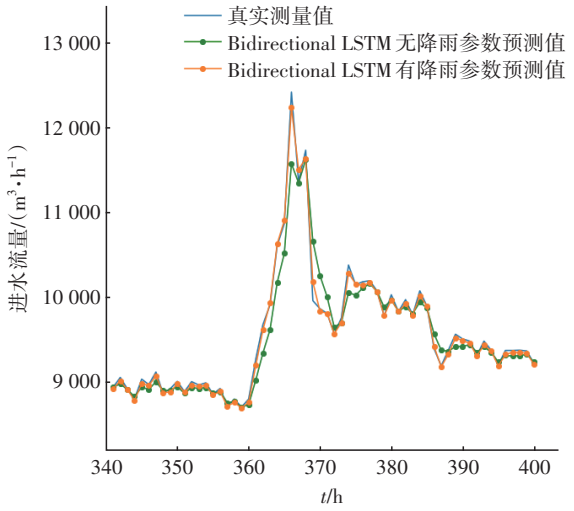


图4　不同方法下进水流量预测结果与真实测量值

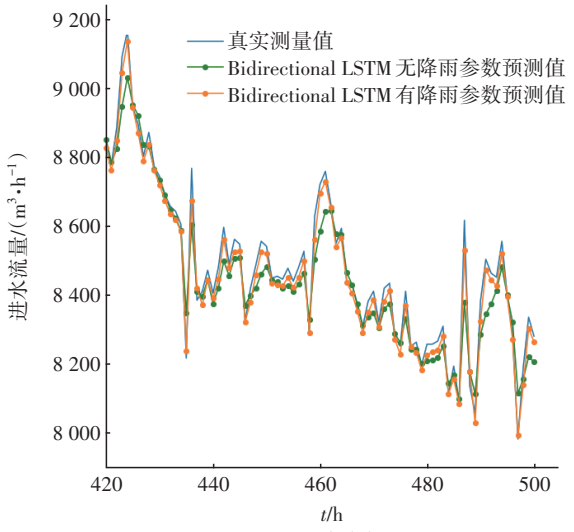
Fig.4　Prediction results and real values of influent flow at different methods

图5对比了有无降雨作为输入变量情况下的实验预测效果。从图5(a)可以看出，加入降雨变量后，Bidirectional LSTM方法的预测精度显著提升，不仅可以较好地预测波峰，还能有效减少模型预测的滞后性。从图5(b)可以看出，日常事件增加降雨变量后预测精度也有显著提升，一是因为增加参数为系统提供了更多信息；二是在此基础上，训练产生了新的、整体精度更高的模型。

为验证模型的普适性，实验还对江苏省宜兴市某污水处理厂进水流量数据进行了测试，图6为有降雨输入下的预测结果（前30个样本为训练集，后续数据为测试集）。可知，Bidirectional LSTM模型具有很好的泛化能力。



a. 降雨事件



b. 日常事件

图 5　有无降雨参数下 Bidirectional LSTM 预测结果

Fig.5　Prediction results of Bidirectional LSTM with or without rainfall

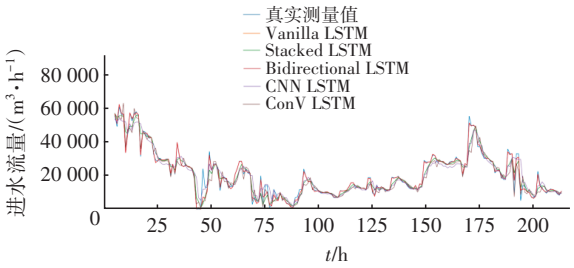


图 6　有降雨情况下进水流量预测结果

Fig.6　Prediction results of inflow flow with rainfall

3　结论

基于 5 种 LSTM 神经网络模型，在有无降水参数下，对江苏省无锡市某污水厂进水流量数据进行了预测，并对其进行测试，取得了很好的效果。本案例中，Bidirectional LSTM 最优的实验参数条件是：LSTM 的隐藏层单元数为 250，训练轮数为 200，训练集样本数为 250。在相同实验设置下，Bidirectional LSTM 模型相比其他方法能够更有效地预测污水厂的进水流量。在双变量预测下，即增加区域内降雨参数后，同等实验参数设置下，相比单变量可以提升近 20% 的流量预测精度。

参考文献：

[1] 苟非洲，程玉婷. 基于长短期记忆网络的日供水量预测方法研究[J]. 中国给水排水，2019，35(17)：79-83. GOU Feizhou, CHENG Yuting. Daily water supply forecasting method based on long short-term memory network [J]. China Water & Wastewater, 2019, 35(17): 79-83(in Chinese).

[2] 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.

[3] 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.

[4] ZHANG D, LINDHOLM G, RATNAWEERA H. Use long short-term memory to enhance Internet of Things for combined sewer overflow monitoring [J]. Journal of Hydrology, 2018, 556: 409-418.

[5] ZHANG D, MARTINEZ N, LINDHOLM G, *et al*. Manage sewer in-line storage control using hydraulic model and recurrent neural network [J]. Water Resources Management, 2018, 32(6): 2079-2098.

[6] KARIMI H S, NATARAJAN B, RAMSEY C L, *et al*. Comparison of learning-based wastewater flow prediction methodologies for smart sewer management [J]. Journal of Hydrology, 2019, 577: 123977.

作者简介：李双宇（1996-　），女，河北保定人，硕士研究生，研究方向为污水处理数学模拟算法。

E-mail：lsy_shirley@pku.edu.cn

收稿日期：2021-06-01

修回日期：2021-07-05

（编辑：任莹莹）