DOI:10.20040/j.cnki.1000-7709.2023.20230137

基于机器学习的洪水预报实时校正

衣学军[1],汤 岭[2],李致家[2],盛奕华[2],姚 成[2],杜若愚[2]

(1. 山东省水文中心,山东 济南 250000; 2. 河海大学水文水资源学院,江苏 南京 210098)

摘要:为了提高临沂流域水文模型的实时洪水预报精度,基于临沂流域的下垫面特征,建立了临沂流域的TOPKAPI网格模型,采用BP神经网络和LSTM模型对TOPKAPI模型模拟结果在不同预见期内进行了校正,在此基础上使用了堆叠方法并选用Transformer模型作为二级学习器,对BP和LSTM的校正结果进行了二次学习。结果表明,经过BP和LSTM模型的实时校正,TOPKAPI模型模拟精度得到了明显提高,预见期越短,校正效果越好;在经过堆叠方法进行二次学习后,校正效果最佳,可有效提升临沂流域洪水预报精度。

关键词:TOPKAPI模型;实时校正;BP神经网络;LSTM模型;洪水预报;临沂流域

中图分类号:TV122;P338 **文献标志码**:A **文章编号**:1000-7709(2023)12-0078-04

1 引言

由于影响因素较多,洪水预报存在难以避免的误差。实时校正有助于提高预测精度,而机器学习可有效地提取数据特征。近年来,机器学习在实时校正方面的应用越来越广泛。姚超宇等[1]使用GBDT方法构建误差校正模型并将其应用于洪水预报,相比基于AR方法和KNN方法的校正模型具有更高的预报精度;汪昊燃等[2]使用KNN算法和反馈法对水位进行实时校正,提高了模拟精度;余宇峰等[3]通过时空图卷积网络构建映射函数并证明其在洪水预报实时校正的适用性;张旭旻等[4]构建基于误差修正思想的自回归模型进行洪水实时校正,校正效果优于传统自回归模型;SUN Y等[5]改进了动态系统响应曲线法的结构,校正后提高了模型输出的精度;张艳等[6]引入动态系统响应校正技术对面雨量进行实时修正处理,提升了新安江洪水预报的效果。然而,目前在实时校正方面的研究主要集中在单一方法上,对于多模型校正结果的融合研究较少;在校正模型选用方面,神经网络和基于注意力机制的模型在实时校正方面的应用较少。因此,本文以临沂流域为例,选取BP和LSTM模型对TOPKAPI模型模拟结果建立校正系统,并将两种模型作为基学习器采用堆叠方法进行校正,二级学习器选用Transformer模型,分析实时校正能力,以提高TOPKAPI模型的洪水预报精度。

2 研究流域和数据来源

临沂水文站以上流域面积为10 315 km^2,流域形状呈扇形,河长287.5 km,属温带大陆性季风气候,年降水量约800 mm。沂河上兴建了大量的水利拦河工程,流域内有4个大型水库,本文将4个水库控制区域产生的径流作为入流处理,流域高程、水系、站点分布及流域边界见图1。

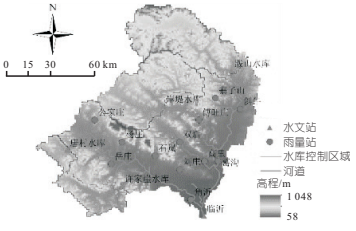

图1 临沂流域概况

Fig. 1 Overview of Linyi Basin

模型输入需要的雨水情数据来源于山东省水文局,数据时间序列为2011~2021年,时间间隔为2 h,气温数据来自中国地面气候资料日值数

收稿日期:2023-02-05,**修回日期**:2023-04-11

基金项目:国家自然科学基金项目(52079035)

作者简介:衣学军(1971-),男,教授级高级工程师,研究方向为水文水资源分析计算、水土保持监测等,E-mail:171290106@qq.com

通讯作者:李致家(1962-),男,博士、教授,研究方向为水文预报与水文模型,E-mail:zhijia-li@vip.sina.com

据集(V3.0),并采用正弦函数插值成 2 h,数字高程数据为 SRTMDEM 产品(90 m)。土壤数据采用世界土壤数据库(Harmonized World Soil Database version 1.1,HWSD),土地利用数据取自中国土地利用现状遥感监测数据库数据集,分辨率均为 1 km。

3　研究方法

3.1　TOPKAPI 模型

TOPKAPI 模型[7]是以模块化思想构建,其中包含了蒸散发、河道径流和地表径流等水文计算模块。模型计算单元为 DEM 格网,气象数据采用距离平方倒数法计算每个格网内的净雨量和蒸散发量。地表径流、地下径流等通过非线性水库描述。具体产流机制见图 2。

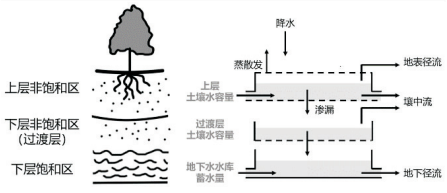

图 2　TOPKAPI 模型产流机制

Fig. 2　TOPKAPI model flow generation mechanism

3.2　神经网络模型

BP 神经网络基于误差反向传播(又称误差反传)算法的多层前馈神经网络进行训练,单隐含层的 BP 网络可模拟任何闭区间内的连续函数。本文选取一层隐含层,隐含层的节点数设置为 30 个。

3.3　LSTM 模型

长短期记忆网络[8](LSTM)为一种时间序列算法,相比于循环神经网络(RNN)能处理梯度消失和梯度爆炸的问题,LSTM 循环结构内部增加了门控结构,即增加了 3 个门(遗忘门、输入门、输出门)。

模型构建时主要涉及到的参数有 LSTM 的神经元个数、批大小的个数、迭代次数、优化器的选择、模型的评估标准等。首先是神经元个数与批大小个数的确定,先选择不同的神经元个数,经过试错法发现当神经元个数为 100 时,模型的损失函数曲线最后趋于平衡,模型较稳定。因为数据量不大,所以设置批大小为全部样本,保证模型的收敛效率,迭代次数设为 300。然后选用 Adam 来训练模型,用均方误差(M_{MSE})来评估模型。

3.4　组合校正方法

采用堆叠方式将 BP 神经网络和 LSTM 模型组合起来,即将 BP 和 LSTM 的预测结果作为新的学习样本进行再一次学习,初始样本的标签仍作为二级学习器的标签,为保证二次学习的效果,二级学习器与一级学习在原理上应具有差异性,因此选用自注意力机制的 Transformer 模型作为二级学习器,旨在通过 3 种完全不同的网络结构去学习误差序列特征,提高校正效果。

Transformer 模型是基于多头注意力机制的模型,模型的主要结构主要是由输入层、位置编码、编码器、解码器及输出层组成,取 Ecoder 和 Decoder 层数均为 6 层,输入向量拓展维度为 512,主要流程见图 3。

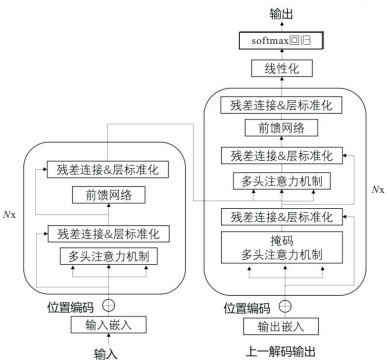

图 3　Transformer 基本结构图

Fig. 3　Transformer basic structure diagram

其中注意力计算公式为:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\boldsymbol{V}\mathrm{softmax}(\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}/\sqrt{d_k}) \tag{1}$$

式中,Attention 为计算出的注意力值;$\boldsymbol{Q}$ 为查询矩阵;$\boldsymbol{K}$ 为键矩阵;$\boldsymbol{V}$ 为值矩阵;softmax 为计算注意力权重的一个函数;d_k 为隐藏层的维度。

3.5　模型评价指标

分别引入洪峰误差 $R_{\mathrm{RE_p}}$、峰现时间误差的绝对值 $T_{\mathrm{TE_p}}$、确定性系数 N_{NSE} 作为评价指标,具体公式为:

$$R_{\mathrm{RE_p}}=|(q_{\mathrm{s,max}}-q_{\mathrm{o,max}})/q_{\mathrm{o,max}}|\times 100\% \tag{2}$$

$$T_{\mathrm{TE_p}}=|T_{\mathrm{s}}-T_{\mathrm{o}}| \tag{3}$$

$$N_{\mathrm{NSE}}=1-\frac{\sum_{i=1}^{n}(q_{\mathrm{s},i}-q_{\mathrm{o},i})^2}{\sum_{i=1}^{n}(q_{\mathrm{o},i}-\bar{q}_{\mathrm{o}})^2} \tag{4}$$

式中,$q_{\mathrm{s,max}}$ 为模拟峰值;$q_{\mathrm{o,max}}$ 为实测峰值;T_{s}、T_{o} 分别为模拟、实测峰现时间,h;$q_{\mathrm{s},i}$、$q_{\mathrm{o},i}$ 分别为系列 i 的模拟值、实测值;$\bar{q}_{\mathrm{o}}$ 为实测均值;n 为场次资料的长度。

4　基于机器学习的洪水预报实时校正

选取临沂流域 2010～2021 年 13 场典型洪

水，采用 TOPKAPI 模型对其进行洪水模拟，模型输入为降雨、温度、土地利用类型和土壤类型等数据，通过 PreTPK 预处理工具箱将值赋予到每个计算网格，因大部分参数可通过下垫面直接估算，所以采用人工优选法对参数进行微调，参数值见表 1，将 13 场洪水中前 9 场作为训练期，后 4 场作为验证期。

表 1 TOPKAPI 模型参数

Tab. 1 Parameters of TOPKAPI model

参数类型	参数名称	参数值
土壤厚度	I-B-2c	0.55
	Lc101-2a	0.86
横向饱和水力传导度	I-B-2c	5.35×10^{-9}
	Af52-3b	7.36×10^{-8}
纵向饱和水力传导度	I-B-2c	5.35×10^{-9}
	Af52-3b	7.36×10^{-8}
河道曼宁系数	一级河道	0.086
	二级河道	0.076
	三级河道	0.065
	四级河道	0.058
	五级河道	0.035
地表曼宁系数	Continuous urban fabric	0.10
	Discontinuous urban fabric	0.16
	Industrial or commercial units	0.01
	Port areas	0.05
	Airports	0.12
	Dump sites	0.05
	Green urban areas	0.06
	Sport and leisure facilities	0.06
	Rice field	0.06
	Fruit trees and berry plantations	0.10

4.1 TOPKAPI 模拟结果分析

表 2 为临沂断面计算洪水的洪峰误差、峰现时间误差和确定性系数的模拟结果。

表 2 TOPKAPI 模型模拟结果

Tab. 2 Simulation results of TOPKAPI model

时期	洪号	评价指标		
		R_{RE_p}/%	T_{TE_p}/2h	N_{NSE}
率定期	2011081800	15.65	7	−0.19
	2011091000	27.16	5	0.58
	2012070700	19.88	2	0.72
	2012072200	42.12	2	0.24
	2013052506	4.24	1	−1.81
	2016080500	70.79	2	−5.30
	2017071312	131.62	2	−7.89
	2018081608	13.12	2	0.75
	2019080308	12.63	1	0.75
	平均值	37.47	2.67	−1.35
验证期	2020080600	21.22	6	0.33
	2020081100	1.99	1	0.87
	2021082600	8.22	8	0.70
	2021092400	13.93	5	0.76
	平均值	11.34	3.33	0.66

由表 2 可知，TOPKAPI 模型在临沂流域模拟的大部分洪水场次表现效果较好，然而，由于部分实测洪水具陡涨陡落的特点，该模型在这种情况下很难与实际观测结果匹配，因此模拟误差相对较大。这可能是由于雨量和流量数据是从水文局的实时库中导出，经过了插值处理，因此实际观测过程线存在偏差，呈涨落较大的趋势。另外，人为因素对模拟精度也有一定的干扰。尽管模型将上游 4 个大型水库的控制面积流量作为河道入流处理，但在流域中仍存在许多中小型水库未纳入模型考虑，这些因素的存在也容易造成系统性误差，且洪峰越小的场次误差越大。因此，提出了一种基于机器学习的实时校正方法，以提高模拟精度并弥补其存在的不足。

4.2 实时校正结果

根据洪水模拟结果与实测结果得到所有场次误差序列，通过机器学习模型分析误差序列之间的关系，分别以 2、6、12 h 为预见期，分析临沂断面实时校正后的洪水模拟改进效果。具体流程如下：①数据划分。将 13 场洪水的误差序列拼接在一起，其中训练集数据为前 9 场，测试集数据为后 4 场。②构造样本。使用滑动窗口的方法构造，时间步长为 10，即用 10 个时段的数据作为 X，后一个预见期的时段作为 Y，由此得到训练样本和测试样本。③样本归一化。为防止测试集的归一化信息影响到模型训练，不再根据序列样本最大、最小值归一化，而采取固定的能稳定包含数据序列的范围段对数据进行归一化。④一层模型训练和验证。将训练样本放入 LSTM 和 BP 模型进行训练和验证，得到各自训练集的 Y_{train} 和验证集的 Y_{test}。⑤二层模型训练和验证。将一层模型得到的 Y_{train} 作为特征，对应时段的实测流量作为标签进行训练，验证时特征使用的是一层模型的验证结果 Y_{test}，标签使用的是对应时段的实测流量。

表 3 为各校正结果的统计。

表 3 各校正模型模拟结果统计

Tab. 3 Statistical of simulation results for each correction model

时期	BP 模型校正			LSTM 模型校正			组合校正模型		
	M_{RE_p}/%	M_{TE_p}/2h	M_{NSE}	M_{RE_p}/%	M_{TE_p}/2h	M_{NSE}	M_{RE_p}/%	M_{TE_p}/2h	M_{NSE}
率定期 2 h	9.88	1.22	0.90	6.79	0.89	0.92	8.67	1.11	0.92
验证期 2 h	5.09	1.50	0.96	5.32	2.00	0.95	3.59	1.25	0.97
率定期 6 h	25.17	2.56	0.46	22.24	1.89	0.51	20.20	1.89	0.32
验证期 6 h	7.77	2.50	0.86	14.28	3.00	0.86	10.40	3.00	0.84
率定期 12 h	33.18	3.56	−0.01	31.29	3.00	−0.83	27.69	2.44	0.13
验证期 12 h	9.09	5.00	0.76	13.49	4.75	0.71	11.37	4.75	0.76

注：M_{RE_p} 为洪峰相对误差的绝对平均值；M_{TE_p} 为峰现时间误差的绝对平均值；M_{NSE} 为平均纳什系数。

由表 3 可看出，对于 BP 和 LSTM，各预见期内确定性系数均较原 TOPKAPI 模型模拟结果有所提升，洪峰相对误差绝大多数情况下有所下降。2 h 预见期下 3 种方法无论是在率定期还是

验证期,平均洪峰误差均在 10%以内,M_{NSE} 均在 0.9 以上,相较之前模拟结果有大幅提高。随着预见期的增加校正效果有所降低,但仍比之前模拟结果好。图 4 为临沂断面 2021092400 号洪水实测流量、模拟流量及 BP、LSTM 和二次学习后各预见期的实时校正过程线。由图 4 可知,2021092400 号实测洪水具陡涨陡落的特点,但由于模型结构限制,原 TOPKAPI 模型模拟结果为光滑曲线,与实测流量过程偏差较大。在 2 h 预见期的校正下,模型模拟精度大大提高,得到的峰形与实测结果几乎一致。当预见期增加时,各模型的曲线出现不同程度的抖动,其中,BP 模型校正在 12 h 预见期的抖动最大,而 LSTM 模型则更平缓,这可能是由于误差序列是通过实测与模拟的差值而来,具有一定的时序性,而 LSTM 模型具有长时记忆功能,因此更平缓。在率定期中,各预见期堆叠方法的校正效果均优于前两种方法,但在验证期中各指标各有优劣,由不同校正方法不同预见期的箱线图(图 5)可看出,在不同预见期的大部分指标中堆叠方法得到结果更集中,说明堆叠方法的校正结果更稳定。总体而言,LSTM 模型相比 BP 模型具有更好的校正结果,而堆叠方法能综合两种模型的优点,从而进一步提高校正精度。

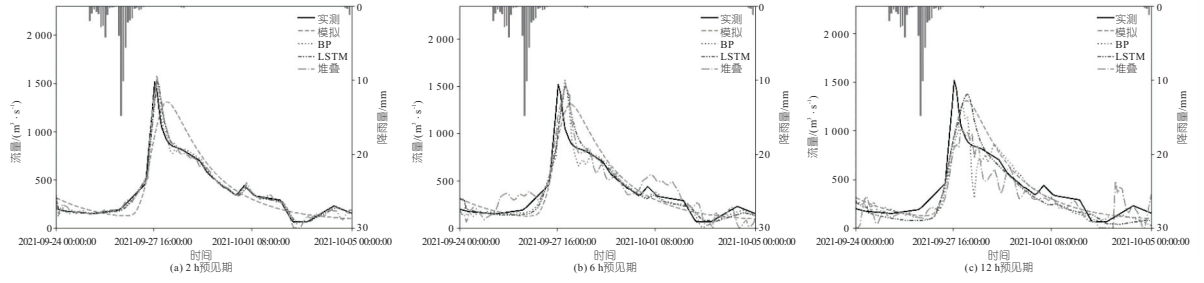

图 4　2021092400 号洪水 2、6、12 h 预见期校正过程线

Fig. 4　No. 2021092400 flood 2 h, 6 h and 12 h forecast period correction process line

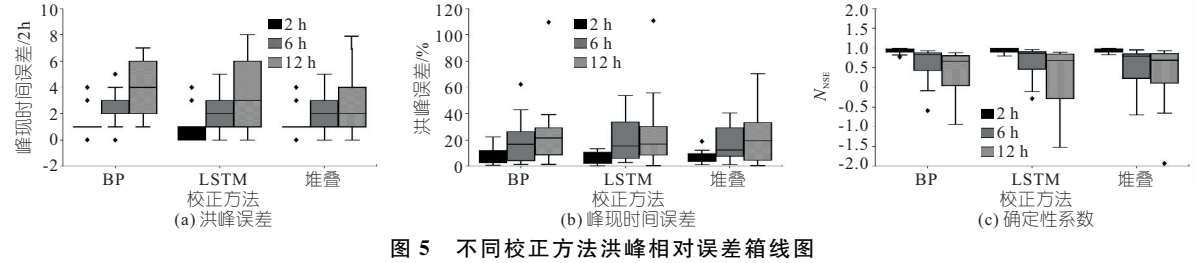

图 5　不同校正方法洪峰相对误差箱线图

Fig. 5　Flood peak relative error box plots of different correction methods

5　结论

a. TOPKAPI 模型充分考虑了研究区域下垫面情况,在临沂流域模拟效果良好,验证了 TOPKAPI 模型的适用性。

b. BP 神经网络和 LSTM 模型均能提高 TOPKAPI 模型的模拟精度,且预见期越短,模拟精度越高,对流域的实时预报有一定的指导作用。

c. 以 Transformer 模型作为第二学习器的堆叠方法的校正结果优于对 BP 和 LSTM 的校正结果,验证了堆叠方法的有效性。

参考文献:

[1] 姚超宇,钟平安,徐斌,等. 基于 GBDT 的实时洪水预报误差校正方法[J]. 水电能源科学,2019,37(8):38-42.

[2] 汪昊燃,王容,黄鹏年,等. 水文水力学结合的秦淮河流域洪水模拟与实时校正研究[J]. 河海大学学报(自然科学版),2022:1-8[2022-11-01]. https://kns.cnki.net/kcms/detail/32.1117.TV.20221101.0950.004.html.

[3] 余宇峰,李薇,李珂,等. 基于 STGCN 的洪水预报误差实时校正方法[J]. 水文,2022,42(5):35-40.

[4] 张旭旻,瞿思敏,李倩,等. 基于协整理论的洪水预报实时校正方法及应用研究[J]. 水资源保护,2022:1-14[2022-02-18]. https://kns.cnki.net/kcms/detail/32.1356.TV.20220218.1147.004.html.

[5] SUN Y,BAO W,JIANG P. Development of multi-variable dynamic system response curve method for real-time flood forecasting correction[J]. Water resources research,2018,54(7):4730-4749.

[6] 张艳,梁忠民,陈在妮,等. 大渡河流域上游洪水预报及实时校正研究[J]. 水力发电,2020,46(5):13-15,21.

[7] 刘家琳,梁忠民,李彬权,等. 基于多模型组合的淮河王家坝断面洪水预报[J]. 水电能源科学,2019,37(8):34-37.

[8] 徐冬梅,王逸阳,王文川. 基于贝叶斯优化算法的长短期记忆神经网络模型年径流预测[J]. 水电能源科学,2022,40(12):42-46.

(下转第 67 页)

报结果在洪水起涨阶段会出现局部低估和高估现象,对退水段的预报结果趋向于低估流量过程,总体上预报能力下降较为明显。其原因可能为在较长预见期时,当前时刻输入的前期流量序列与输出的预报流量之间的关联性下降,导致区域化 LSTM 洪水预报模型的预报能力下降。

4　结论

a. 通过合成水文一致区内各流域场次洪水资料,建立区域化 LSTM 洪水预报模型,为乏资料流域洪水预报模型的构建提供了一条有效途径。

b. 在南四湖湖西平原区的示例应用表明,在 15 h 预见期内(即$\leqslant 5\Delta t$),区域化洪水预报模型具有较高的预报精度,当预见期$>$15 h 时,模型的预报精度有所降低。

c. 本研究初步显示了区域化 LSTM 洪水预报模型在乏资料地区的应用潜力,但在资料前处理中仅考虑了流域面积因素的影响,而流域坡度、形状等属性均会影响洪水过程,如何考虑多种流域属性进行区域化建模,有待深入研究。另外,受资料条件限制,本文仅采用了近年 40 场洪水数据,结论具有一定局限性。

参考文献:

[1] 毛能君, 夏军, 张利平, 等. 参数区域化在乏资料地区水文预报中应用研究综述[J]. 中国农村水利水电, 2016(12): 88-92.

[2] 谈戈, 夏军, 李新. 无资料地区水文预报研究的方法与出路[J]. 冰川冻土, 2004,26(2): 192-196.

[3] JIANG S, ZHENG Y, SOLOMATINE D. Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning [J]. Geophysical research letters, 2020, 47(13):e2020GL088229.

[4] 殷兆凯, 廖卫红, 王若佳, 等. 基于长短时记忆神经网络(LSTM)的降雨径流模拟及预报[J]. 南水北调与水利科技, 2019, 17(6): 1-9, 27.

[5] 陶思铭, 梁忠民, 陈在妮, 等. 长短期记忆网络在中长期径流预报中的应用[J]. 武汉大学学报(工学版), 2021, 54(1): 21-27.

Research on LSTM-based Regionalized Flood Forecasting Model

BI Cheng-lin[1], LIU Kuang[2], XIANG Zheng[2], WANG Jun[1], QIAN Ming-kai[3], LIANG Zhong-min[1]

(1. College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China;
2. Hydrology Center of Shandong Province, Jinan 250000, China;
3. Hydrology Bureau of the Huaihe Water Conservancy Commission, Bengbu 233001, China)

Abstract: Limited by hydrometeorological data, flood forecasting in ungauged basins still faces challenges. Parameter regionalization is a common method to solve this problem. The machine learning model has the characteristics of simple modeling and convenient use compared with the traditional flood forecasting model. Taking the West Plain of Nansihu Lake in Shandong Province as the research area, referencing the idea of hydrological regional synthesis, this paper synthesizes the data of 40 floods in 8 watersheds from 2010 to 2021, and builds a regionalized flood forecasting model based on Long Short-Term Memory (LSTM). The results show that the regionalized flood forecasting model can simulate the actual flood process well, the relative error of flood peak in both the training set and the testing set are less than 10%, and the Nash-Sutcliffe efficiency coefficients are all greater than 0.9; In the 15 h forecast period, the regionalized flood forecasting model has higher forecasting accuracy, and when the forecast period is more than 15 h, the forecast accuracy of the model decreases.

Key words: regionalized models; flood forecasting; ungauged basins; LSTM

(上接第 81 页)

Real-time Correction of Flood Forecasting Based on Machine Learning

YI Xue-jun[1], TANG Ling[2], LI Zhi-jia[2], SHENG Yi-hua[2], YAO Cheng[2], DU Ruo-yu[2]

(1. Shandong Provincial Hydrological Center, Jinan 250000, China;
2. College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China)

Abstract: In order to enhance the real-time flood forecasting accuracy in the Linyi River Basin, a TOPKAPI grid model was developed based on the underlying surface characteristics of the Linyi River Basin. The TOPKAPI model simulation results were corrected at different lead times using BP neural networks and LSTM models. Furthermore, a stacking approach was applied, employing the Transformer model as a secondary learning tool to refine the corrections made by BP and LSTM. The results indicate that after real-time correction with the BP and LSTM models, the improvement of the simulation accuracy of the TOPKAPI model is obvious, with better correction results for shorter lead times. Following the stacking method for secondary learning, the correction results is the best, effectively enhancing the flood forecasting accuracy in the Linyi River Basin.

Key words: TOPKAPI model; read-time correction; BP neural network; LSTM model; flood forecasting; Linyi River Basin