

DOI: 10. 20040/j. cnki. 1000-7709. 2023. 20231253

洪水预报区域化 LSTM 模型研究

毕成琳¹, 刘 匡², 向 征², 王 军¹, 钱名开³, 梁忠民¹

(1. 河海大学水文水资源学院, 江苏 南京 210098; 2. 山东省水文中心, 山东 济南 250000;
3. 淮水利委员会水文局, 安徽 蚌埠 233001)

摘要: 受水文气象资料条件限制, 在乏资料地区开展洪水预报仍面临挑战。虽然水文模型参数区域化是解决该问题的常用途径, 但机器学习模型因其具有建模简单、使用方便等特点, 已逐渐成为构建乏资料地区洪水预报方案的重要方法。以山东省南四湖湖西平原区为例, 借鉴水文地区综合的思想, 对 2010~2021 年 8 个流域共 40 场暴雨洪水资料进行合成, 构建了基于长短期记忆神经网络(LSTM)的区域化洪水预报模型。研究表明, 区域化 LSTM 洪水预报模型能够较好地模拟实际洪水过程, 训练集和测试集的洪峰流量相对误差均小于 10%, 纳什效率系数均大于 0.9; 在 15 h 预见期内, 区域化洪水预报模型具有较高的预报精度, 当预见期超过 15 h 时, 模型的预报精度有所降低。

关键词: 区域化模型; 洪水预报; 乏资料地区; LSTM

中图分类号: [TV124]; P333

文献标志码: A

文章编号: 1000-7709(2023)12-0063-05

1 概况

山东省南四湖湖西平原区(以下简称“湖西平原区”)位于鲁西南地区, 属暖温带半湿润季风区大陆性气候, 暴雨是造成本区域洪水的主要因素。该区域虽建有水文观测站点(图 1), 但站点相对稀疏、场次雨洪数据有限, 属乏资料地区。加之该区域地处黄泛平原, 地势总体上低洼平缓, 易积涝成灾, 是典型的平原型暴雨径流一致区。目前, 实现乏资料地区洪水预报最常用的方法是参数区域

化^[1], 其核心思想是将有资料地区(参证流域)率定的水文模型参数移植到乏资料地区(目标流域)^[2], 或将小尺度的模型参数推广至大尺度。这类方法的精度主要取决于所选用参证流域的代表性, 并与使用的水文模型和区域化方法有关。近年来, 随着大数据分析技术的快速发展, 利用深度学习的超强数据挖掘和建模能力进行水文过程的模拟预测已成为研究热点^[3-4]。相较于高度依赖水文要素及物理机制的传统水文模型, 深度学习模型无需精细刻画水文过程中间环节, 避开了难以获取的水文状态变量, 通过时序数据直接训练预报变量与影响因素之间的关系, 可较容易建立预报方案。其中, 长短期记忆神经网络(LSTM)因其在处理时序问题中的表现力, 在水文预报中具有显著应用优势^[5]。为此, 本文借鉴水文地区综合的思想, 在水文一致区内进行各流域场次洪水资料合成, 构建了基于 LSTM 的区域化洪水预报模型, 评估了区域化模型的适用性, 为乏资料流域洪水预报提供了参考。



图 1 研究区位置及主要测站分布

Fig. 1 Location of the study area and the distribution of main stations

2 研究思路与方法

收稿日期: 2023-07-27, 修回日期: 2023-08-28

基金项目: 水利部重大科技项目(SK2022032); 国家自然科学基金项目(41730750)

作者简介: 毕成琳(1999-), 女, 硕士研究生, 研究方向为水文水资源, E-mail: bichenglin1999@163.com

通讯作者: 梁忠民(1962-), 男, 博士、教授、博导, 研究方向为水文水资源, E-mail: zmliang@hhu.edu.cn

2.1 研究思路

对于同一个水文一致区的各个流域,认为其形成暴雨的天气系统相似、自然地理属性类同,所以各流域暴雨径流规律较为一致。一致区内包含众多流域,其中一些流域具有一定数量的场次暴雨洪水资料,而有些流域没有或缺乏次洪资料。对于单个乏资料流域而言,由于缺乏或可用的次洪资料有限,无法建立本流域洪水预报 LSTM 模型,但可通过收集该流域所在水文一致区内其他有资料流域的场次洪水资料,构成区域场次洪水大数据;依赖 LSTM 神经网络的时间回归预测能力,构建区域统一的洪水预报模型,不仅可实现区域内乏资料流域的洪水预报,对有资料流域,因为利用了一致区内其他流域的更多样本信息,亦可改进本流域的预报效果。

2.2 研究方法

2.2.1 LSTM 模型原理

LSTM 模型结构见图 2。LSTM 网络的核心在于单元状态的更新和传递,遗忘门从单元状态中丢弃非重要信息,输入门决定单元状态中需要更新的信息,输出门决定需要输出的信息。图 2 中,σ 表示 Sigmoid 激活函数,tanh 表示 tanh 激活函数,设置激活函数的作用在于更新或遗忘信息,激活函数示意图见图 3。

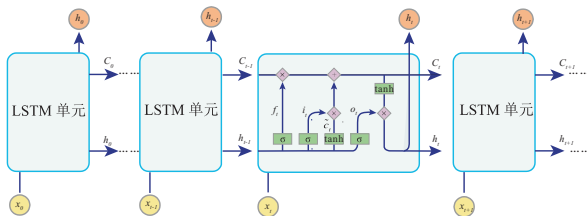


图 2 LSTM 模型结构图

Fig. 2 Structure of the LSTM model

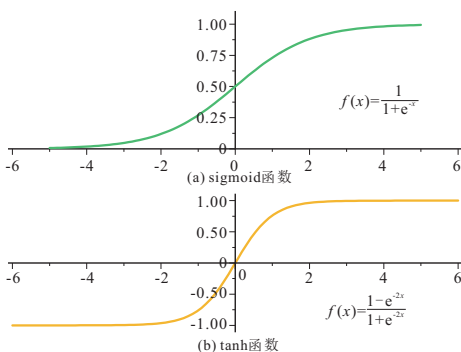


图 3 激活函数示意图

Fig. 3 Diagram of the activation function

图 2 中,当前 t 时刻的输入为 x_t ,内部控制门结构包括遗忘门、输入门和输出门,其中遗忘门为:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

输入门为:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

单元状态:

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (4)$$

输出门为:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t \times \tanh(C_t) \quad (6)$$

式中, f_t, i_t, o_t 分别为 t 时刻的遗忘门、输入门、输出门的输出向量; W, b 分别为权重矩阵和偏移向量,下标表示具体到哪个门; h_t, h_{t-1} 分别为 $t, t-1$ 时刻的隐藏状态; C_t, C_{t-1} 分别为 $t, t-1$ 时刻的单元状态更新向量; \tilde{C}_t 为 t 时刻的单元状态候选向量。

2.2.2 基于 LSTM 的区域化洪水预报模型构建

(1)模型前处理与样本划分。认为一致区内的暴雨天气系统和自然地理属性相似,径流形成机制较为一致,加之空间位置相近,因此可以将一致区内各流域有限场次洪水数据集合在一起构建数据驱动模型。由于区域内各流域面积大小不一,而流域面积作为最主要的流域特征,降雨一定是决定流域洪水大小的最重要因素。当将不同流域的洪水数据合并建模时,为消除流域面积大小的影响,首先对原始流量数据做模数化处理:

$$\tilde{Q}_k(t) = Q_k(t) / F_k \quad (7)$$

式中, $\tilde{Q}_k(t)$ 为 k 流域 t 时刻的流量模数; $Q_k(t)$ 为 k 流域 t 时刻原始流量数据; F_k 为 k 流域的面积。

由于 8 个流域的实测洪水数据间隔为 2、6 h,为满足研究区实际预报需要,将数据时间间隔通过线性内插方式统一处理为 3 h,共计得到 6 517 组样本。将样本划分为训练集和测试集,以数据资料较为充分、场次洪水数较多的孙庄、鱼台、张庄闸和路菜园闸 4 个流域的 31 场洪水作为模型训练集,对区域化 LSTM 洪水预报模型进行训练;以场次洪水数缺乏的后营、马庄闸、李庙闸和黄寺闸 4 个流域的 9 场洪水作为模型测试集,评价区域化 LSTM 洪水预报模型的性能。场次洪水编号与预报断面水文站的对应关系见表 1。

表 1 场次洪水编号与预报断面水文站对应表

Tab. 2 Correspondence table between flood numbers and hydrological stations at the outlet of the basin

| 预报断面 | 孙庄 | 鱼台 | 张庄闸 | 路菜园 | 后营 | 马庄闸 | 李庙闸 | 黄寺闸 |
|--------|-----|------|-------|-------|-------|-----|-------|-------|
| 水文站 | | | | 园闸 | | | | |
| 场次洪水编号 | 1~5 | 6~23 | 24~27 | 28~31 | 32~33 | 34 | 35~37 | 38~40 |

(2)区域化 LSTM 洪水预报模型的结构。合理选取主要影响因子是建立 LSTM 模型的关键。

考虑到资料的有限性和预报方案对乏资料流域的可用性,选择预报断面前期流量及断面以上前断面平均雨量作为预报因子,输出变量为预报断面不同预见期的流量值。通过综合分析湖西平原区各流域的降雨历时和平均汇流时间,最终确定以前期 12 个时段长($\Delta t = 3 \text{ h}$)共 36 h 作为预报因子的输入长度,构建了未来 8 个预见期(1~8) Δt 的洪水预报模型(图 4)。本文重点在于构建区域化洪水预报模型,不讨论降雨预报精度对模型结果的影响,因此在模型的训练及测试中,以实测降雨序列作为预见期降雨的输入。

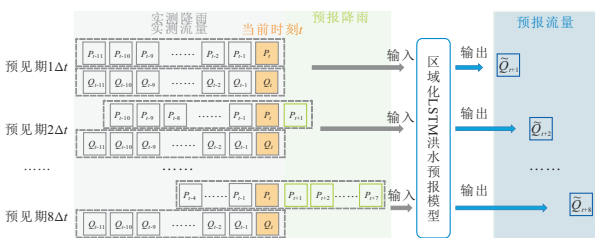


图 4 模型输入输出示意图

Fig. 4 Diagram of the model input and output

本次构建的 LSTM 神经网络,隐藏层层数为 2 层,每层隐藏层的神经元数量分别设置为 20、10,输出层为包含 1 个神经元的全连接输出层,即每次输出一个结果。训练模型时,学习率设定为 0.1,每次从训练集中随机选取 512 个样本,最大训练次数取 3 000。选择 Adam 算法作为模型优化算法,并选取平均绝对误差(M_{MAE})作为损失函数。 M_{MAE} 的计算公式为:

$$M_{MAE} = \frac{1}{n} \sum_{i=1}^n |y_{true}^i - y_{pred}^i| \quad (8)$$

式中, y_{true}^i 为第 i 时刻的真实值; y_{pred}^i 为第 i 时刻的模型预测值; n 为时间步长总数。

由于输入输出变量的量纲不一致,为加快 LSTM 神经网络的学习速度和收敛速度,采用 MAX-MIN 对模型输入数据集进行归一化处理:

$$\hat{X}_i = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad 1 \leq i \leq N \quad (9)$$

式中, \hat{X}_i 为归一化处理后的第 i 个元素; X_i 为原始数据序列中的第 i 个元素, X_{max} 、 X_{min} 分别为原始序列中的最大、最小值; N 为序列长度。

2.3 研究数据

收集一致区内 8 个流域 2010~2021 年间共 40 场暴雨洪水数据(数据来源于山东省水文中心),各流域相关信息见表 2。

3 结果与分析

3.1 模型精度评价指标

选取洪峰流量相对误差(R_{REP})、洪量相对误

表 2 流域基本信息

Tab. 2 Basic information of the watersheds

| 预报断面 水文站 | 所属河流 | 集水面 积/ km^2 | 雨量 站数 | 水文 站数 | 场次 洪水数 | 数据时间 间隔/h |
|-------------|-------|-------------------------|----------|----------|-----------|--------------|
| 后营 | 梁济运河 | 3 225 | 5 | 1 | 2 | 6 |
| 孙庄 | 万福河 | 1 199 | 5 | 1 | 5 | 2 |
| 鱼台 | 东鱼河 | 5 998 | 10 | 1 | 18 | 2 |
| 张庄闸 | 东鱼河 | 3 934 | 4 | 3 | 4 | 6 |
| 马庄闸 | 东鱼河北支 | 755 | 3 | 1 | 1 | 6 |
| 路菜园闸 | 东鱼河 | 646 | 3 | 1 | 4 | 6 |
| 李庙闸 | 东鱼河南支 | 938 | 1 | 1 | 3 | 6 |
| 黄寺闸 | 胜利河 | 3 225 | 3 | 1 | 3 | 6 |

差(R_{REV})、峰现时间绝对误差(E_{EPT})、纳什效率系数(N_{NSE})作为模型精度评价指标。采用合格率(Q_{QR})指标评定区域化洪水预报模型的整体预报精度。当一次预报的误差小于许可误差时,为合格预报。本文分别以实测洪峰流量的 20%、一个时段长(3 h)和 0.7 作为洪峰流量、峰现时间和 N_{NSE} 的许可误差。合格预报次数与预报总次数之比为合格率。计算公式为:

$$Q_{QR} = (n/m) \times 100\% \quad (10)$$

式中, n 为预报合格次数; m 为预报总次数。

3.2 结果与分析

(1)模型整体精度分析。区域化 LSTM 洪水预报模型训练集及测试集的精度统计结果见表 3。其中,场次洪水 1~31、32~40 分别用于区域化模型的训练和测试。由表 3 可知,训练集 31 场洪水的洪峰流量相对误差均在 $\pm 10\%$ 以内,绝对值平均为 2.96%;洪量相对误差较小,绝对值平

表 3 训练集及测试集的精度统计结果

Tab. 3 Accuracy results of floods for the training set and the testing set

| 场次 洪水 编号 | 洪峰流量 相对误 差/% | 洪水总 量相对 误差/% | 峰现时间 绝对误 差/h | N_{NSE} | 场次 洪水 编号 | 洪峰流量 相对误 差/% | 洪水总 量相对 误差/% | 峰现时间 绝对误 差/h | N_{NSE} |
|----------------|--------------------|--------------------|--------------------|-----------|----------------|--------------------|--------------------|--------------------|-----------|
| 1 | -1.03 | -1.15 | 6 | 0.98 | 21 | 0.24 | -0.27 | -6 | 0.98 |
| 2 | 2.32 | 0.59 | -3 | 0.98 | 22 | 0.82 | 0.36 | -3 | 0.98 |
| 3 | 4.32 | -0.34 | 0 | 0.97 | 23 | -3.51 | 5.88 | 0 | 0.94 |
| 4 | -0.97 | -1.22 | 0 | 0.99 | 24 | -0.64 | -0.53 | 0 | 0.93 |
| 5 | 3.02 | 0.30 | 0 | 0.96 | 25 | 1.62 | 0.77 | -3 | 0.97 |
| 6 | 9.64 | -1.33 | 3 | 0.93 | 26 | 7.78 | 0.92 | 3 | 0.95 |
| 7 | 2.60 | -0.57 | 3 | 0.97 | 27 | -2.17 | -1.75 | 3 | 0.97 |
| 8 | 3.69 | 0.26 | 0 | 0.96 | 28 | 2.74 | -0.54 | 3 | 0.94 |
| 9 | -1.56 | -0.41 | 3 | 0.96 | 29 | 2.84 | 1.77 | 0 | 0.96 |
| 10 | -3.90 | -0.13 | 3 | 0.82 | 30 | -9.98 | -5.07 | 3 | 0.95 |
| 11 | 1.34 | -0.33 | -3 | 0.98 | 31 | -8.31 | -4.66 | 3 | 0.97 |
| 12 | 1.72 | 0.69 | 0 | 0.97 | 32 | -0.72 | -0.92 | 3 | 0.99 |
| 13 | -1.53 | 4.28 | 0 | 0.98 | 33 | 12.56 | 0.35 | -3 | 0.97 |
| 14 | -0.39 | -1.32 | 3 | 0.97 | 34 | -8.30 | -5.16 | 0 | 0.91 |
| 15 | 2.40 | 0.52 | 3 | 0.84 | 35 | 6.93 | -0.64 | 0 | 0.97 |
| 16 | 0.54 | -0.89 | 3 | 0.86 | 36 | 5.98 | -0.68 | 0 | 0.96 |
| 17 | -2.30 | -1.87 | 3 | 0.99 | 37 | 9.81 | -0.73 | 0 | 0.94 |
| 18 | -6.49 | -1.63 | 0 | 0.66 | 38 | 1.65 | -1.76 | 3 | 0.96 |
| 19 | -1.11 | 0.11 | 3 | 0.99 | 39 | -3.55 | -1.37 | 3 | 0.93 |
| 20 | -0.29 | -1.05 | 3 | 0.97 | 40 | -3.66 | -1.50 | 0 | 0.98 |

均为 1.34%；峰现时间绝对误差均在 ±6 h 以内，绝对值平均为 2.23 h；31 场洪水的平均 N_{NSE} 为 0.94；训练集预报合格的场次共 29 场，合格率为 93.5%。

测试集 9 场洪水的洪峰流量相对误差除第 33 号洪水外，其余均在 ±10% 以内，绝对值平均为 5.91%；洪量相对误差较小，绝对值平均为 1.46%；峰现时间绝对误差均在 1 个时段以内，绝对值平均为 1.33 h；9 场洪水的 N_{NSE} 均大于 0.9，平均 N_{NSE} 为 0.96；测试集预报合格场次共 9 场，合格率为 100%。图 5 为测试集 2 场洪水(场次 33、38)的模拟过程线。由图 5 可知，区域化 LSTM 洪水预报模型能够较好地反映出洪水过程的趋势，对洪峰流量、峰现时间的预测与实测较为接近，整体看模型预测趋势能较好地拟合实测流量的变化过程。

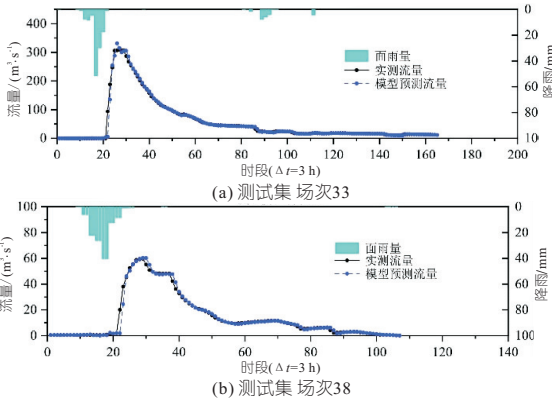


图 5 测试集场次洪水模拟过程线

Fig. 5 Flood event simulation process line for the testing set

(2)模型精度随预见期变化分析。为考查不同预见期下区域化 LSTM 模型的预报能力，绘制不同预见期下 4 项评价指标的箱线图，见图 6，对应数值结果见表 4。可以看出，对计算的 8 个预见期(24 h)，区域化 LSTM 模型对洪峰、洪量均具有较高的预报能力；当预见期超过 $5\Delta t$ (15 h) 后，模型精度有所降低，其中 N_{NSE} 指标开始低于 0.7，降速相对较快。

(3)典型洪水预报效果分析。以测试集场次编号 40 洪水为例，绘制不同预见期下的区域化 LSTM 模型预报流量过程和实测流量过程，见图 7。由图 7 可知，区域化 LSTM 洪水预报模型在洪水预报方面效果较好。具体而言，由图 7(a)~(e)可知，在预见期为 (1~5) Δt 时，区域化 LSTM 洪水预报模型对洪水过程的预报与实际情况较为接近，对洪峰流量的预报能力也较好，在退水段也与实际过程较为吻合。当预见期超过 $5\Delta t$ 后(图 7(f)~(g))，区域化 LSTM 模型的预

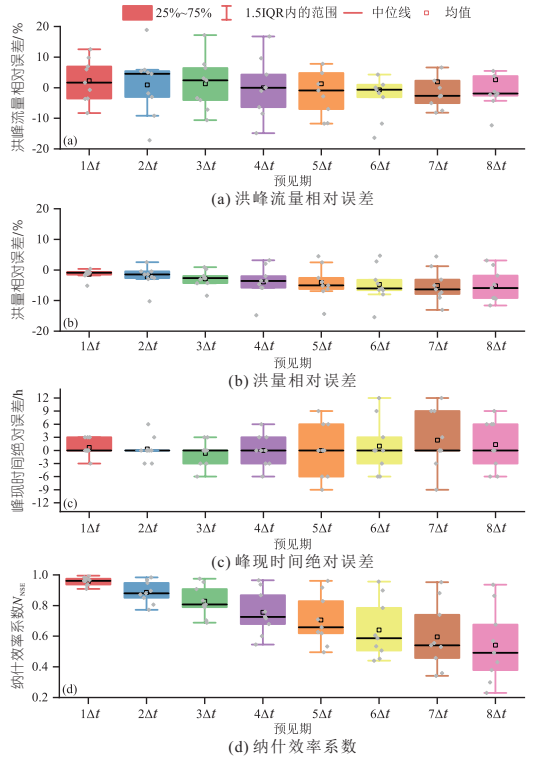


图 6 不同预见期下的预报结果评价指标箱线图

Fig. 6 Box graph of evaluation indicators of forecast results under different forecast periods

表 4 测试集不同预见期的模型统计结果

Tab. 4 Statistical results of the model for different forecast periods on the testing set

| 预见期 (Δt) | 洪峰流量相 | 洪量相对误 | 峰现时间绝 | N_{NSE} 平均值 |
|-----------------------|-------|-------|-------|------------------|
| | 对误差绝对 | 差绝对值平 | 对误差绝对 | |
| | 值平均/% | 均/% | 值平均/h | |
| 1 | 5.91 | 1.46 | 1.33 | 0.96 |
| 2 | 7.87 | 2.73 | 1.66 | 0.89 |
| 3 | 6.88 | 3.18 | 2.00 | 0.83 |
| 4 | 6.84 | 5.05 | 2.67 | 0.76 |
| 5 | 8.89 | 5.68 | 4.67 | 0.71 |
| 6 | 6.76 | 6.41 | 4.33 | 0.64 |
| 7 | 7.74 | 6.38 | 5.00 | 0.60 |
| 8 | 8.08 | 6.27 | 4.67 | 0.54 |

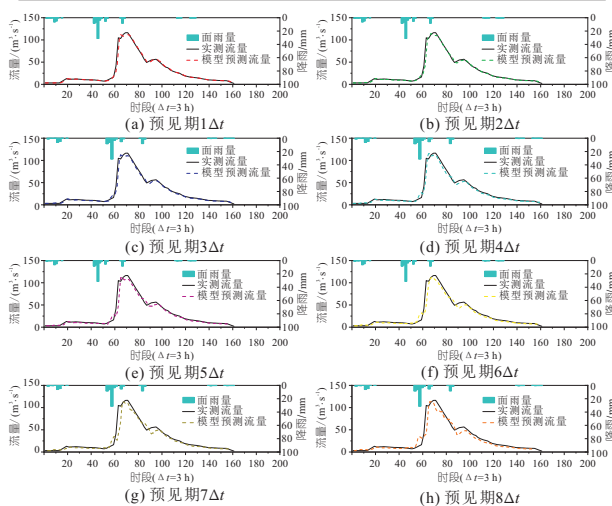


图 7 场次编号 40 洪水不同预见期下的模拟结果

Fig. 7 Simulation results of flood number 40 under different forecast periods

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 $\leq 5\Delta t$),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-lin¹, LIU Kuang², XIANG Zheng², WANG Jun¹, QIAN Ming-kai³, LIANG Zhong-min¹

(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-jun¹, TANG Ling², LI Zhi-jia², SHENG Yi-hua², YAO Cheng², DU Ruo-yu²

(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