DOI： 10.14042/j.cnki.32.1309.2025.01.005

考虑遥相关因子的月降水数据偏差校正方法

闫宝伟[1,2]，常建波[1,2]，孙明博[1,2]，古东霖[1,2]，周学叡[1,2]，杨东旭[1,2]

（1. 华中科技大学土木与水利工程学院，湖北 武汉，430074；2. 华中科技大学数字流域科学与技术湖北省重点实验室，湖北 武汉，430074）

摘要：多源降水数据校正对于缺资料地区水文规律的分析及模拟至关重要，当前校正方法对气候要素考虑不足。为此，基于 ERA5、ERA5-Land、MSWEP-V2 和 PERSIANN-CDR 多源降水数据集与遥相关因子集，结合 XGBoost-SHAP 模型进行特征筛选与成因分析，构建基于 BiLSTM 的降水数据偏差校正模型，采用贝叶斯优化（BO）策略寻求模型的最优超参数组合，以进一步提高校正精度。选取汉江上游为研究对象，对多源降水数据进行偏差校正。结果表明：①大气环流类因子是汉江上游降水形成的主要影响因素，北半球副高脊线位置指数的影响最大；②与传统的统计类方法相比，BO-BiLSTM 略逊色于表现最优的参数转换法，但可以更加灵活地考虑多个因子的影响；③考虑遥相关因子后，多源降水数据校正的测试期纳什效率系数平均提升了 5.4%，均方误差平均降低了 24.6%，Kling-Gupta 效率系数平均提升了 10.5%。研究成果可为数据匮乏地区月降水的高精度估算与延长提供切实可行的技术方案。

关键词：月降水；偏差校正；遥相关因子；多源降水数据集；BiLSTM 模型

中图分类号：P407; P333　　**文献标志码：**A　　**文章编号：**1001-6791（2025）01-0050-12

月降水序列对于水文模拟、径流预报和水资源评价具有重要的应用价值[1-3]，特别是缺资料地区的水文预测中，月降水序列作为最为关键的输入，其质量直接决定了水文模型的精度[4]。当前中国虽然已布设了大量雨量站网，但由于各站网之间投入使用时间不一，高一致性的月降水序列仍然较短[5]。近年来，卫星技术与数值模拟技术快速发展，卫星产品与再分析数据集的出现可缓解月降水的序列长度与一致性问题。

当前各类反演产品与再分析产品层出不穷，极大地丰富了降水数据资源。降水反演产品通过对遥感数据进行反演得到，如 MSWEP-V2 数据集[6]、PERSIANN 数据集[7] 和 GPM 类数据集[8]。其中，GSMaP 作为 GPM 的重要产品，提供了更高时空分辨率的降水数据[9]。再分析产品则是深度融合了动力学模型模拟结果与观测数据，利用一定规则所生成的具有空间均匀性和时间连续性的降水产品，如欧洲中期天气预报中心（ECMWF）利用预报模型和同化系统对多源观测数据进行再分析而生成的 ERA5 与 ERA5-Land 数据集[10]。这 2 类产品时间序列长、覆盖范围广，可在一定程度上反映区域降水的变化趋势，但其数值与实际降水数据之间仍然存在一定偏差，不能直接用于工程计算与水文模拟[11]。若结合实际数据对其进行偏差校正，充分发挥不同降水数据之间的互补性，则可得到一套高精度、长序列且近乎实况的月降水系列。

常见的降水校正方法有线性回归、分布函数映射和贝叶斯方法等[12]，它们均可在一定程度上校正降水数据，但这类方法依赖于预设的数学或统计模型，可能与实际情况不符[13]，机器学习模型则不需要太多假设，模型复杂性和结构可根据数据自适应。随机森林（RF）、极端梯度提升模型（XGBoost）由于其计算效率和拟合能力最先被广泛使用，包括降水数据的偏差校正与降尺度研究[14-15]。近年来，随着人工智能技术的发展，长短期记忆网络（LSTM）、卷积神经网络（CNN）等层数更多、结构更复杂的深度学习模型也逐渐用于降水的偏差校正，结果均表明该类模型对非线性时序数据处理具有更好的性能[16]。

收稿日期：2024-10-12；**网络出版日期：**2025-02-08

网络出版地址：https://link.cnki.net/urlid/32.1309.P.20250208.1106.002

基金项目：国家重点研发计划项目（2021YFC3200301）；国家自然科学基金项目（52079054）

作者简介：闫宝伟（1981—），男，山东滨州人，副教授，博士，主要从事水文基础理论研究。E-mail：bwyan@hust.edu.cn

通信作者：常建波，E-mail：changjianbo_water@163.com

然而，以上校正方法一般考虑待校正数据集与环境变量。当降水为小时或者日尺度时，以上变量足以辅助降水的偏差校正；但当降水为月尺度数据时，上述变量则略显不足。遥相关因子一定程度上反映了大气的运动与状态，对全球及区域气候具有重要影响。若引入这些因子辅助降水数据的校正，某种程度上可以考虑降水的形成机制，从而提高校正精度。沙普利加法解释模型(SHAP)可以量化每个特征因子的贡献程度，常被用来解释机器学习模型的预测结果，目前还未有研究利用该方法进行融合遥相关因子的降水数据校正分析。

本研究采用 XGBoost-SHAP 筛选降水的驱动要素，基于贝叶斯优化的双向 LSTM(BO-BiLSTM)构建多源降水数据偏差校正模型，对各类数据集的月降水进行校正，并在汉江上游进行应用研究。

1　研究区域与数据

1.1　研究区域

汉江上游流域面积近 9.5 万 km^2，是南水北调中线工程水源地和国家一级水源保护区。流域地貌类型复杂，水系发达，水资源丰富，降水量充沛，整个区域属于亚热带季风气候，降水空间与年内分布不均，年际变化较大。图 1 为该区域地形和雨量站点概况图。

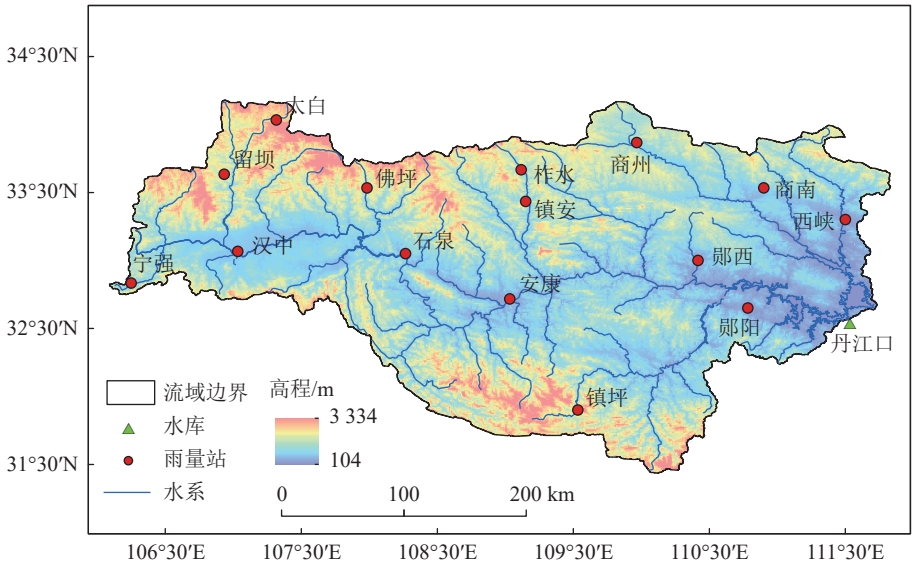


图 1　汉江上游流域概况

Fig. 1　Overview of Upper Hanjiang River basin

1.2　数　据

(1)地面观测数据。汉江上游降水数据来自国家气象信息中心(http://data.cma.cn)，包括流域内 15 个气象站点 1971—2020 年的逐日降水数据，通过对日尺度实测降水进行累加得到月降水数据，采用泰森多边形进行面积加权得到逐月面降水量，所选用站点的分布见图 1。

(2)多源降水数据集。本研究拟结合多个降水数据集验证所提出方法的性能，主要包括 ERA5、ERA5-Land、MSWEP-V2 和 PERSIANN-CDR 数据集。其中，ERA5、ERA5-Land 数据集是 ECMWF 最新和最先进的全球再分析产品，MSWEP-V2 数据集是美国普林斯顿大学所研制的全球首个高分辨率长时间降水数据集[17]，PERSIANN-CDR 数据集是基于人工神经网络算法提取的多卫星降雨数据集[18]。以上数据集涵盖再分析产品与卫星反演产品，各数据集详细信息见表 1。以图 1 矢量区为边界，提取各数据集汉江上游月尺度面降水量。

表 1 多源降水数据集基本信息

Table 1 Basic information of multi-source precipitation datasets

数据集名称	分辨率	时间范围	来源
ERA5-Land	0.1°×0.1°	1971-01/2020-12	https://cds.climate.copernicus.eu
ERA5	0.25°× 0.25°	1979-01/2019-12	https://cds.climate.copernicus.eu
MSWEP-V2	0.1°× 0.1°	1979-02/2020-12	http://gloh2o.org/
PERSIANN-CDR	0.25°× 0.25°	1983-01/2020-12	http://chrsdata.eng.uci.edu/

(3)遥相关因子数据集。来源于国家气候中心的百项气候系统指数集(http://cmdp.ncc-cma.net/Monitoring/cn_index_130.php)，涵盖了大气、海洋、陆地、冰冻圈等各个子系统的监测指数，反映了气候系统内部及其相互作用的关键特征。该指数集中存在部分数据缺失情况，为减少由此带来的误差，将严重缺测的指数删除，并对少量缺测的数据选用 KNN(K-Nearest Neighbors)算法进行插值处理[19]。此外，该类因子对降水的影响存在一定的时间滞后，因此需要对输入的遥相关因子进行时滞处理[20]。时滞设置为 1 ~ 12 个月，共生成遥相关因子特征 1 368 项，全部特征因子以编号形式呈现。

2 研究方法

2.1 降水偏差校正方法

本研究提出一种考虑遥相关因子的降水数据偏差校正方法(图 2)，其主要程序为：

(1)以流域实际降水为基准，训练 XGBoost 模型，得到最优模型后将其与训练期数据输入 SHAP 模型，可完成对遥相关因子集的贡献度排序，进而确定贡献最大的前 n 个特征因子，分析这些特征因子对降水的驱动机制。

(2)将所得的 n 个特征因子与待校正降水数据共 n+1 个因子输入 BO-BiLSTM 模型进行训练和测试，分析降水数据校正结果。

由于遥相关因子需进行 1 a 的时滞处理，故遥相关因子的时间序列比其他数据集多 1 a，所有多源降水数据集训练期、验证期和测试期的划分比例均为 6∶2∶2，统计成果时，将训练期与验证期合并为新的训练集进行分析。

2.2 基于 XGBoost-SHAP 模型的特征筛选

XGBoost 模型[21]是大规模并行决策树运行工具，利用梯度下降框架来提升弱学习器，以实现梯度的高效提升，与其他同类型的算法相比，XGBoost 能在更短的时间内实现更优的分类、回归功能。利用网格搜索法对 XGBoost 模型超参数树的数量(n)、最大树深(d)和学习率(r)进行调优，最终确定 n=100，d=4，r=0.3。

本文选用 XGBoost 模型构建汉江上游面降水的特征筛选模型，将已提取的多源降水数据集的面降水与上述 1 368 项遥相关因子特征一并输入模型。由于这些因子与多源降水数据存在一定的信息冗余，这些冗余的特征会使模型精度下降，计算时间增长。因此，需要对这些因子完成进一步筛选。

SHAP 模型是一种通过计算 Shapley 值、用于解释机器学习计算结果的加性解释模型，对于树模型具有更高的效率与更好的解释性[22]。该模型将所有的特征都视为贡献者，计算其贡献度，贡献度可能为正值，也可能为负值，其中正值对预测结果产生正向驱动作用，负值产生负向驱动作用。贡献度的绝对值越大，表明该特征对降水的影响越大。采用 SHAP 模型计算各遥相关因子特征的贡献度，并选取排序前 n 的因子特征作为后续降水数据偏差校正模型的输入因子。

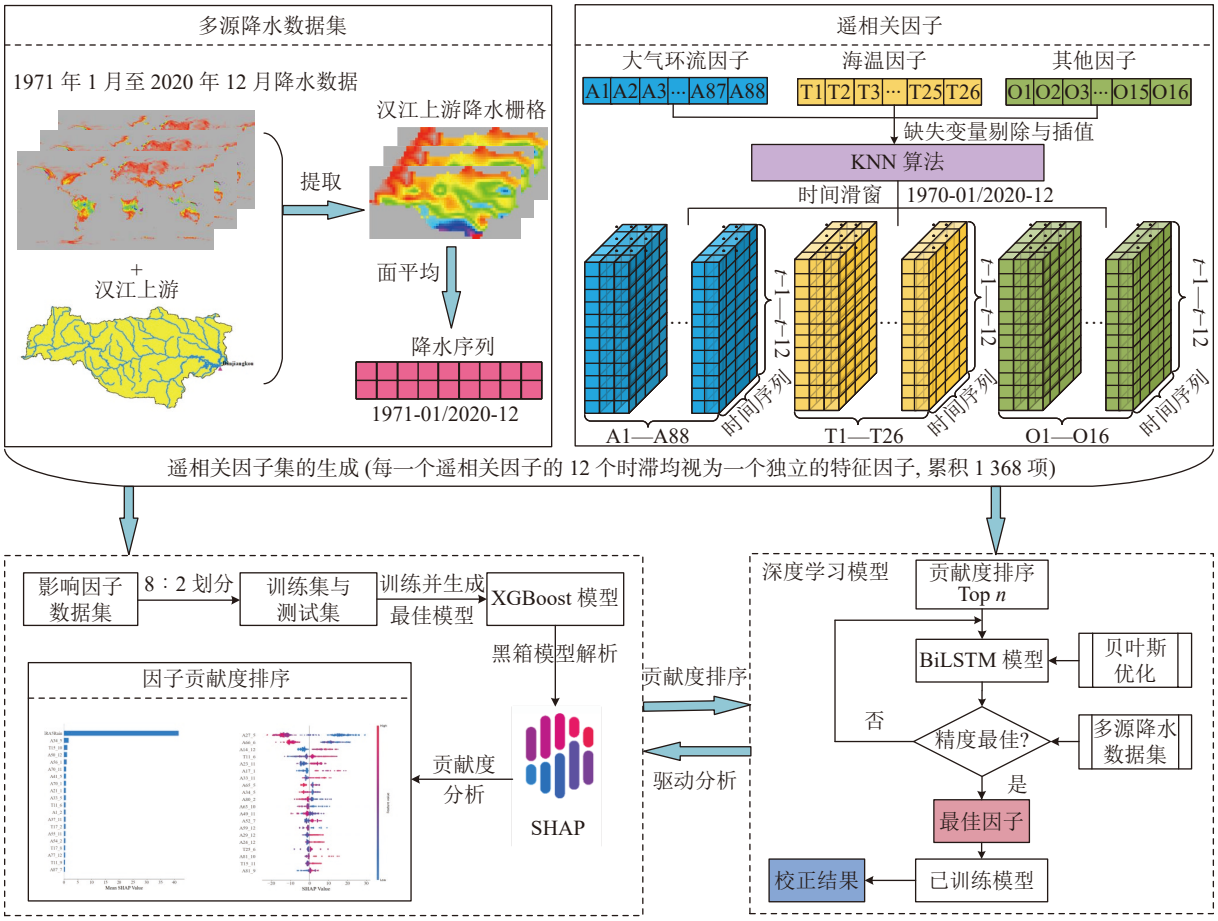


图 2　多源降水数据偏差校正方法示意

Fig. 2　Schematic of bias correction of multi-source precipitation data

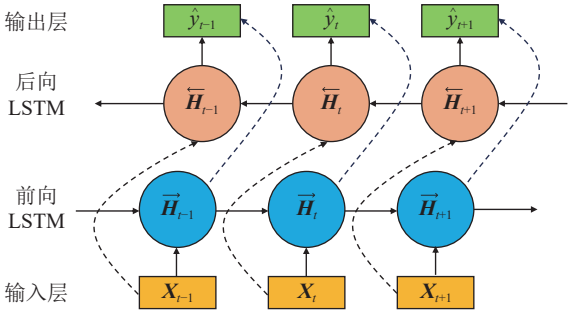
2.3　基于 BO-BiLSTM 的偏差校正模型

常见的偏差校正方法有累积分布函数、统计学校正和深度学习[12,16,23]，其中深度学习具有极强的非线性关系拟合能力，以 LSTM 为代表的模型受到了广泛的认可。为此，选取 BiLSTM 进行降水数据的偏差校正。BiLSTM 模型由前向 LSTM 和后向 LSTM 组成，相较于传统 LSTM 模型，该模型可以更好地获得时序数据的全局特征，并能够充分比较前向数据和后向数据的关联性(图 3)。

将上述筛选的前 n 个特征与多源降水数据作为模型输入，设置当前时刻实际降水为模型输出。考虑到超参数对 BiLSTM 模型性能有较大影响，采用贝叶斯优化(Bayesian optimization，BO)选取模型的超参数组合。以降水校正值与实测值的均方误差为贝叶斯优化的目标函数，选取重点超参数(初始学习率、BiLSTM 神经元数、训练迭代次数)进行优化，并引入 L2 正则项对整个模型进行控制，防止过拟合，运行平台为 MATLAB，所有方案均基于双 NVIDIA-4090D 48G GPU 完成。待优化参数取值范围如表 2 所示[24]。

2.4　对比方案与精度评定

为充分检验本文所提方法的优越性，以当前广泛使用的线性缩放法、参数转换法、分布映射法[12]为基准，设置 5 个对比方案。其中，S1—S3 分别为上述 3 个方法；S4 采用 BO-BiLSTM 模型，但不考虑遥相关因子；S5 则在 S4 基础上增加了筛选出的遥相关因子。同时，选取纳什效率系数(E_{NS})、均方根误差(E_{RMS})和 Kling-Gupta 效率系数(E_{KG})作为精度评价指标[2]。



注：$\boldsymbol{X}_t$ 为 t 时刻输入；$\overrightarrow{\boldsymbol{H}}_t$ 为 t 时刻前向权重向量；$\hat{y}_t$ 为 t 时刻输出；$\overleftarrow{\boldsymbol{H}}_t$ 为 t 时刻后向权重向量。

图 3 BiLSTM 模型结构

Fig. 3 Structure of the BiLSTM model

表 2 待优化参数取值范围

Table 2 Value ranges of parameters to be optimized

参数名称	参数含义	取值范围
r_1	初始学习率	[0.001,0.1]
N	LSTM神经元个数	[1,100]
L_2	L2正则项	$[10^{-10},10^{-2}]$
M	训练最大迭代次数	[1,100]

3 结果及分析

3.1 特征因子的选取

以 ERA5-Land 降水数据为例，采用 XGBoost-SHAP 模型计算各因子的贡献度，按贡献度从高到低依次增加特征因子并将其与 ERA5-Land 数据组合输入 BO-BiLSTM 模型中，观察不同因子组合下实测降水与校正降水的训练期 E_{NS}，以此确定最佳遥相关因子数 n，确定过程可见图 4。由图 4 可知，当遥相关因子数量为 18 时，训练期整体 E_{NS} 最高，为 0.94，若继续增加因子，对于降水数据偏差校正的结果影响较小，故本研究确定最佳输入因子数 n=18，此时 BO-BiLSTM 模型各参数分别为 r_1=0.012，N=6，L_2=5.89×10^{-9}，M=34。

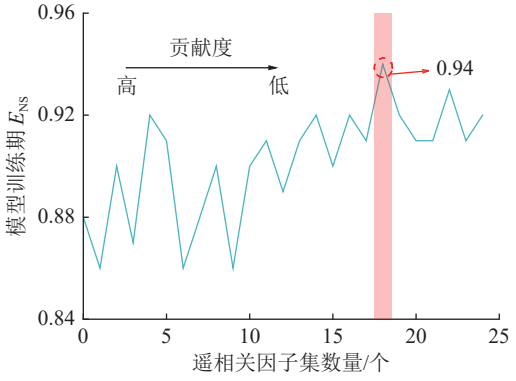


图 4 遥相关最佳输入因子 n 确定过程

Fig. 4 Process of determining the optimal input factor n for teleconnected analysis

整理上述所得前 18 个贡献较大的特征因子，主要含义见表 3。图 5 给出了这些因子特征的平均 SHAP 绝对值及散点密度图。图 5 中从上至下因子特征的重要性逐渐减小，散点颜色映射代表特征因子的值由小到大，每一个点代表一个样本的 SHAP 值，该值代表了这个特征对单个预测的贡献，而点的集合表示了特征整体对预测结果影响的方向和大小。所选取的特征因子中大气环流因子占总数的 67%，表明大气环流因子对于汉江上游降水的驱动贡献强于海温因子和其他因子。

结合表 3 与图 5，A23_5 和 A23_10 对汉江上游降水的贡献度最高，主要表现为负向驱动作用，该指数与大范围环流形势的调整有密切联系。该指数异常偏高时，中国大部分地区以气温偏低为主，长江流域及其以北地区降水偏少[25]。其次为 A66_12，主要表现为正向驱动作用，该指数主要表现青藏高原地区的位势场，通过调节北大西洋、印度洋对青藏高原地区的水汽输送进而影响亚洲地区降水的空间分布与强度[26]。A37_5、A16_1、A29_10 和 A38_5 对汉江上游降水分别产生正向、正向、双向和正向驱动作用，该类副高指数均与大范围环流的形势调整密不可分[25]。A72_1 对汉江上游降水产生正向驱动作用，该指数通过调节区域的位势高度和反气旋性环流对气温产生作用，进而影响中国的区域降水情况。A38_5 和 T6_8 对汉江上游降水产生双向驱动作用，且驱动方式以分段形式呈现，这与厄尔尼诺事件发生后引起大气环流异常密不可分。

表 3　降水校正因子信息

Table 3　Information of precipitation correction factors

编号	物理含义	时间滑窗	所属类别
A23_5	北半球副高脊线位置指数	5	大气环流因子
A66_12	西藏高原-2 指数	12	大气环流因子
A37_5	印度副高北界位置指数	5	大气环流因子
A72_1	东大西洋遥相关型指数	1	大气环流因子
A16_1	西太平洋副高强度指数	1	大气环流因子
A29_10	北美副高脊线位置指数	10	大气环流因子
A38_5	西太平洋副高北界位置指数	5	大气环流因子
T8_5	NINO B区海表温度距平指数	5	海温因子
T24_7	热带印度洋全区一致海温模态指数	7	海温因子
A48_1	北美区极涡面积指数	1	大气环流因子
T16_11	西太平洋暖池强度指数	11	海温因子
T26_9	副热带南印度洋偶极子指数	9	海温因子
T6_8	NINO C区海表温度距平指数	8	海温因子
A23_10	北半球副高脊线位置指数	10	大气环流因子
A85_2	850 hPa西太平洋信风指数	2	大气环流因子
A56_6	北半球极涡中心经向位置指数	6	大气环流因子
T10_9	热带北大西洋海温指数	9	海温因子
A52_5	太平洋区极涡强度指数	5	大气环流因子

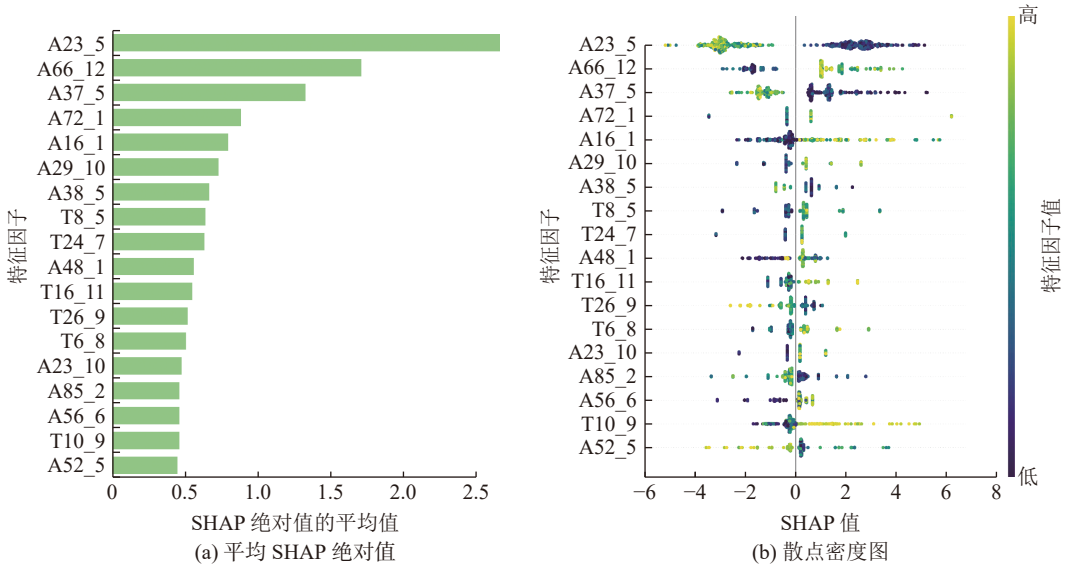


图 5　特征因子筛选结果

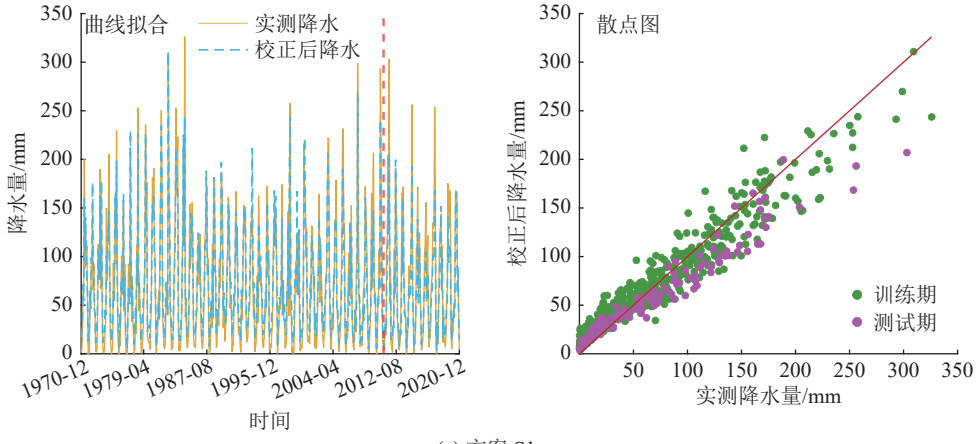
Fig. 5　Results of Feature Factor Screening

T24_7 对汉江上游降水主要产生负向驱动作用，该指数则是通过影响中国以江淮地区为中心的江淮型高温进而作用于降水[27]。A48_1、A56_6 和 A52_5 对汉江上游降水均表现为负向驱动，该类极涡指数通过与南亚高压进行协同，调整东亚大气环流的配置，进而对长江流域的降水产生影响[28]。T16_11 对汉江上游降水主要以正向驱动为主，该指数异常偏高时，亚洲热低压减弱，西太平洋副热带高压加强，位置偏西，进一步使得 850 hPa 风场上中国东部地区为偏北风距平，东亚夏季风减弱，最终使长江流域部分地区降水偏多。

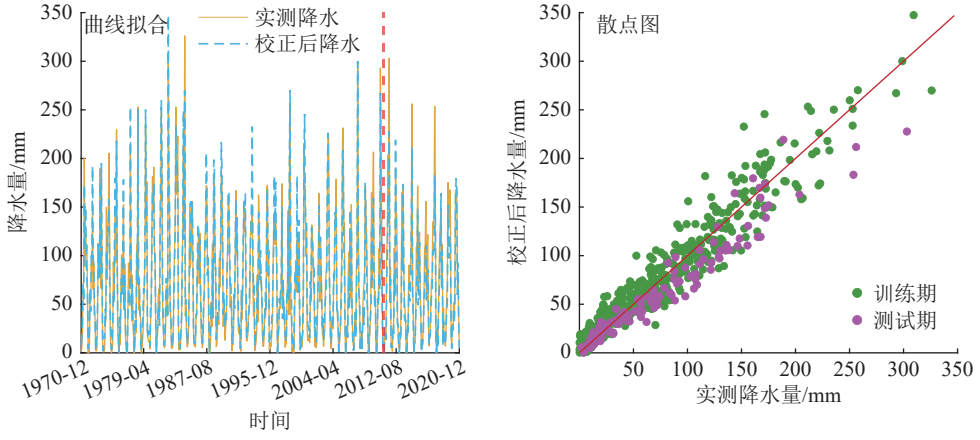
T26_9 对汉江流域上游降水主要表现为负向驱动作用，该指数与印度季风关系密切，可通过影响印度季风进而影响中国的区域降水。A85_2 对汉江流域上游降水主要表现为负向驱动作用，该指数作用于夏季中国降水的 3 条主要雨带，进而影响汉江上游区域降水。T10_9 对汉江流域上游降水主要表现为正向驱动作用，该指数异常时所激发的中纬度波列使得西路冷空气随着西风带槽脊东移，进而影响中国东部降水[29]。

3.2 不同降水校正方法分析

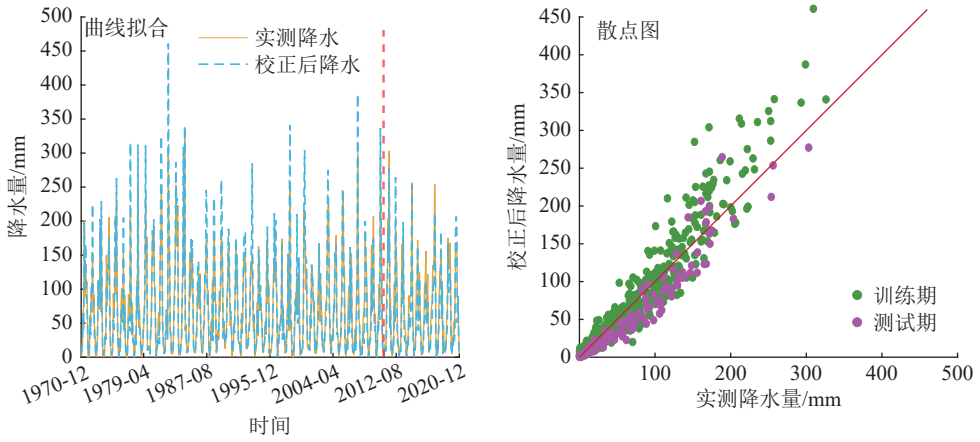
结合筛选后的特征因子，采用 S1—S5 共 5 种设计方案，分别对 ERA5-Land 的再分析降水进行偏差校正，校正结果见图 6(图中红色虚线用于划分训练期与测试期)。可以看出，3 种传统的降水校正方法 S1、



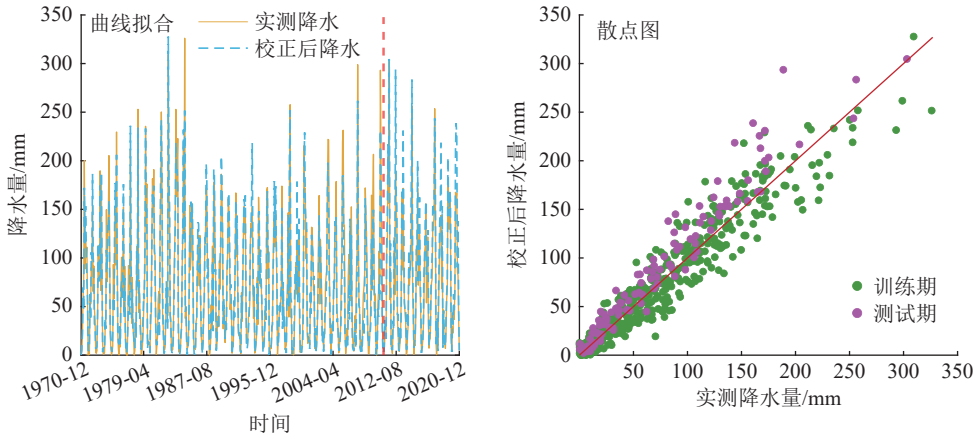
(a) 方案 S1



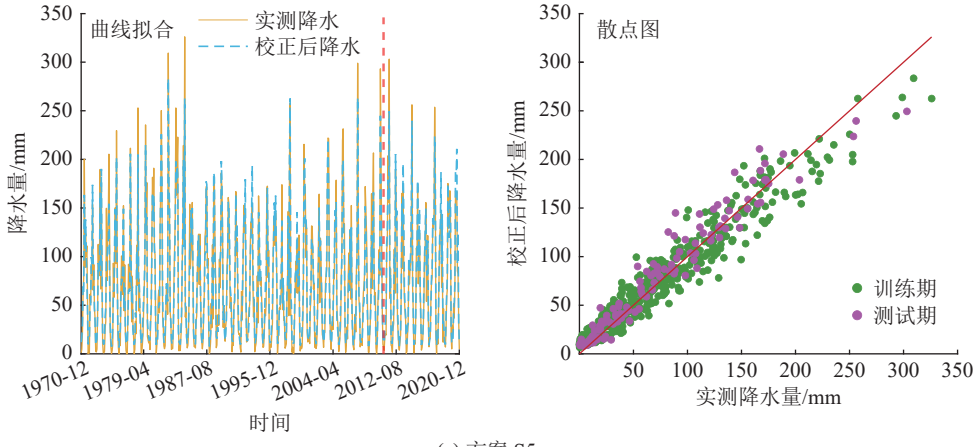
(b) 方案 S2



(c) 方案 S3



(d) 方案 S4



(e) 方案 S5

图 6　各方案校正降水与实测降水拟合情况

Fig. 6　Fitting of calibrated precipitation and measured precipitation in each scheme

S2 和 S3 训练期的校正结果正常，但测试期的校正结果比实际明显偏小，其主要原因为：这些统计学方法预先假定了降水数据的分布，认为数据的趋势是平稳的；S4 与上述情况类似，训练期拟合良好，但测试期计算结果明显偏大，表明不考虑遥相关因子时，深度学习模型捕捉到了降水增加的趋势，但存在显著高估。相比较而言，S5 的校正降水与实际降水曲线拟合良好，散点均匀分布在对角线附近，表明考虑遥相关因子可以有效改善深度学习模型对降水的校正效果。

表 4 进一步给出了各方案训练期与测试期的精度评价结果。当不考虑遥相关因子时，S2 在训练期的效果最好，其次是 S4 和 S1，S3 最差。在测试期，也是 S2 最优，S4 和 S3 次之，S1 最差。BO-BiLSTM 模型虽然略逊于参数转换法，但却可以更灵活地考虑多个因子。进一步考虑遥相关因子后，S5 相对于 S4 的各精度

表 4　各方案降水校正精度

Table 4　Precision of precipitation correction for each scheme

方案	训练期			测试期		
	E_{NS}	E_{RMS} /mm	E_{KG}	E_{NS}	E_{RMS} /mm	E_{KG}
S1	0.91	18.70	0.84	0.86	23.88	0.70
S2	0.93	17.36	0.94	0.89	20.99	0.78
S3	0.84	25.26	0.78	0.87	22.95	0.82
S4	0.92	17.16	0.91	0.88	22.24	0.79
S5	0.94	15.84	0.89	0.94	15.83	0.91

指标均有了较大程度提升，其测试期 E_{NS} 提升了 6.8%，E_{RMS} 降低了 28.8%，E_{KG} 提升了 15.2%。整体来看，S5 相对于这 4 种方法，其测试期 E_{NS} 提升了 5.6%，E_{RMS} 降低了 24.6%，E_{KG} 提升了 11.0%。

3.3 其他降水数据集的应用

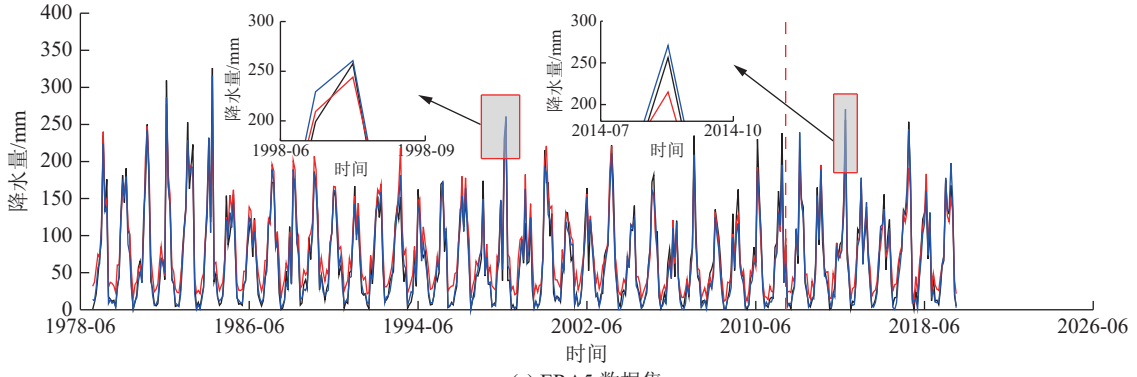
将本研究所提出方法应用于其他降水数据集，校正结果可见表 5 与图 7。由表 5 可知，考虑遥相关因子后，选取的 4 类降水数据集的 E_{NS} 平均提升了 5.4%，E_{RMS} 平均降低了 24.6%，E_{KG} 平均提升了 10.5%。由此可知，本研究所提出的方法对于其他降水数据集具有较好的适用性。进一步地，观察图 7，可以发现，考虑遥相关因子后，校正降水与实际降水在高值与低值处拟合更优，如 1998 年 7 月与 2014 年 8 月的高值降水以及贯穿全时段的低值降水。特别是 MSWEP-V2 降水数据集，未考虑遥相关因子时，各峰值校正结果明显偏大，考虑后得到了显著改善。以上结果进一步表明，考虑遥相关因子可有效改善多源降水数据集中极值的校正效果。

表 5 其他降水数据集校正精度

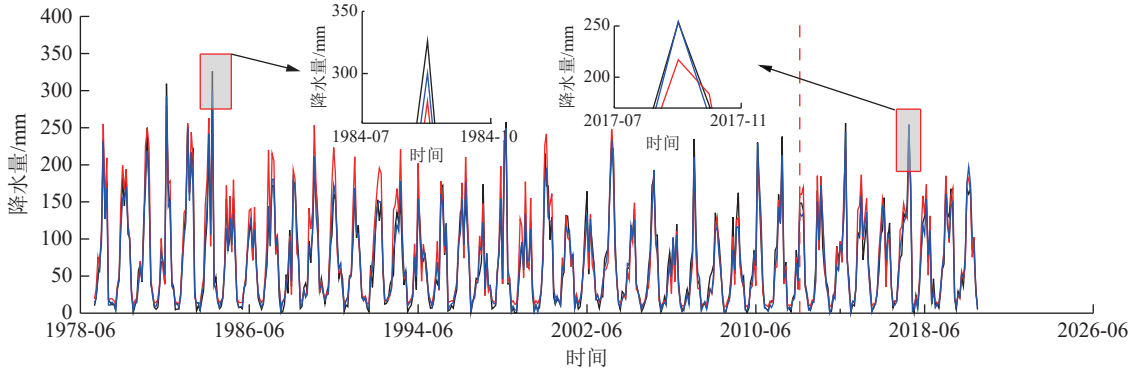
Table 5 Precision of precipitation correction for other precipitation data

数据集及处理	训练期			测试期		
	E_{NS}	E_{RMS} /mm	E_{KG}	E_{NS}	E_{RMS} /mm	E_{KG}
ERA5	0.89	20.87	0.79	0.92	17.15	0.82
ERA5*	0.96	12.83	0.94	0.96	12.48	0.94
MSWEP-V2	0.88	21.51	0.88	0.91	17.46	0.91
MSWEP-V2*	0.95	13.68	0.94	0.95	12.83	0.98
PERSIANN-CDR	0.95	13.20	0.96	0.82	25.24	0.88
PERSIANN-CDR*	0.96	12.59	0.96	0.87	21.21	0.92

注：*表示校正该数据集时输入遥相关因子，反之则无。



(a) ERA5 数据集



(b) MSWEP-V2 数据集

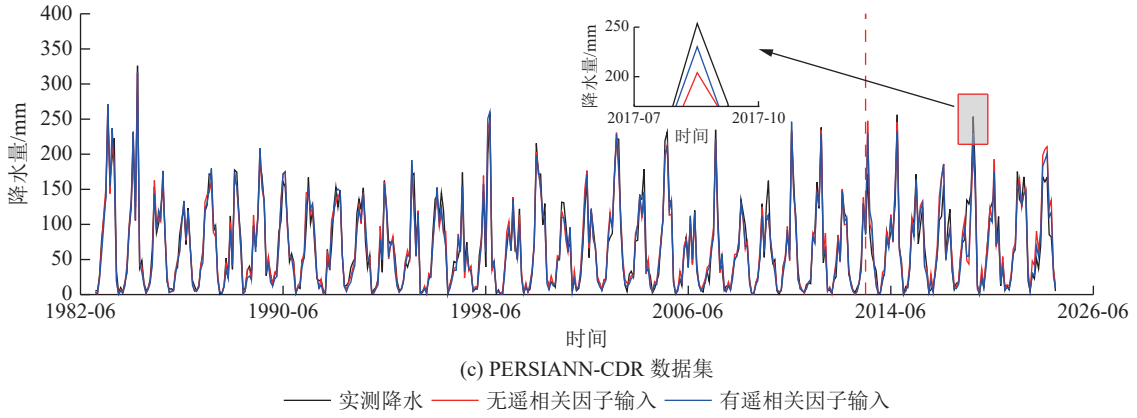


图 7　其他降水数据集的校正结果

Fig. 7　Plots of correction results for other precipitation datasets

4　结　　论

为有效应对降水数据偏差校正过程中气候影响要素考量不足及成因机制解析不深入的问题，本研究以汉江上游为例，基于多源月降水数据与遥相关因子集，采用 XGBoost-SHAP 模型进行了月降水影响因子的筛选及成因剖析，进一步基于 BO-BiLSTM 构建了降水数据偏差校正的深度学习模型。主要结论如下：

（1）大气环流类指数对汉江上游的降水贡献强于海温因子与其他因子，其中，北半球副高脊线位置指数、西藏高原-2 指数和印度副高北界位置指数等贡献相对较高，其影响时间分别为 5 个月、12 个月和 5 个月。

（2）与传统偏差校正技术相比，BO-BiLSTM 的校正精度略低于数据转换法，优于线性缩放法和分布函数法，但 BO-BiLSTM 可以更灵活地考虑遥相关因子的影响。

（3）在 4 个降水数据集中，所述方法具有较强的适用性。与不考虑遥相关因子相比，其训练期纳什效率系数均在 0.9 以上，测试期纳什效率系数平均提升了 5.4%，均方误差平均降低了 24.6%，Kling-Gupta 效率系数平均提升了 10.5%。

本研究所提方法有效提高了汉江上游多个数据集的降水数据偏差校正精度，可为该地区月径流模拟和预报及水资源评价提供重要的数据支撑。今后将继续分析该方法在其他区域的适用性，并进一步探索该方法在栅格尺度上的降水或其他水文气象要素的偏差校正研究，以提供更高精度的水文气象数据产品。

参考文献：

[1] ZOU Y X，YAN B W，GU D L，et al. A water-energy complementary model for monthly runoff simulation[J]. Journal of Hydrology，2024，639：131624.

[2] 孙明博，闫宝伟，常建波，等. 基于物理机制和深度学习的混合模型及应用研究 [J]. 中国农村水利水电，2024(8)：67-72，80.（SUN M B，YAN B W，CHANG J B，et al. A hybrid deep learning model coupled with physical mechanism and its application[J]. China Rural Water and Hydropower，2024(8)：67-72，80.（in Chinese））

[3] 常建波，朱博文，闫宝伟. 黄河流域产水量时空变化及归因分析 [J]. 环境科学与技术，2024，47(8)：113-122.（CHANG J B，ZHU B W，YAN B W. Temporal and spatial variation of water yield in the Yellow River basin and its attribution analysis[J]. Environmental Science & Technology，2024，47(8)：113-122.（in Chinese））

[4] 雍斌，张建云，王国庆. 黄河源区水文预报的关键科学问题 [J]. 水科学进展，2023，34(2)：159-171.（YONG B，ZHANG J Y，WANG G Q. Key scientific issues of hydrological forecast in the headwater area of Yellow River[J]. Advances in Water Science，2023，34(2)：159-171.（in Chinese））

[5] 熊立华，刘成凯，陈石磊，等. 遥感降水资料后处理研究综述 [J]. 水科学进展，2021，32(4)：627-637.（XIONG L H，LIU C

K，CHEN S L，et al. Review of post-processing research for remote-sensing precipitation products[J]. Advances in Water Science，2021，32(4)：627-637.(in Chinese))

[6] BECK H E，van DIJK A I J M，LARRAONDO P R，et al. MSWX：global 3-hourly 0.1° bias-corrected meteorological data including near-real-time updates and forecast ensembles[J]. Bulletin of the American Meteorological Society，2022，103(3)：E710-E732.

[7] ASHOURI H，HSU K L，SOROOSHIAN S，et al. PERSIANN-CDR：daily precipitation climate data record from multisatellite observations for hydrological and climate studies[J]. Bulletin of the American Meteorological Society，2015，96(1)：69-83.

[8] HOU A Y，KAKAR R K，NEECK S，et al. The global precipitation measurement mission[J]. Bulletin of the American Meteorological Society，2014，95(5)：701-722.

[9] YONG B，LIU D，GOURLEY J J，et al. Global view of real-time trmm multisatellite precipitation analysis：implications for its successor global precipitation measurement mission[J]. Bulletin of the American Meteorological Society，2015，96(2)：283-296.

[10] HERSBACH H，BELL B，BERRISFORD P，et al. The ERA5 global reanalysis[J]. Quarterly Journal of the Royal Meteorological Society，2019，146：1999-2049.

[11] 赵君，刘雨，徐进超，等. 基于贝叶斯三角帽法的多源降水数据融合分析及应用 [J]. 水科学进展，2023，34(5)：685-696.(ZHAO J，LIU Y，XU J C，et al. Multi-source precipitation data fusion analysis and application based on Bayesian-Three Cornered Hat method[J]. Advances in Water Science，2023，34(5)：685-696.(in Chinese))

[12] 杜懿，林泽群，王大刚. 不同降水产品在长江流域的偏差校正研究 [J]. 水文，2023，43(1)：62-67.(DU Y，LIN Z Q，WANG D G. Bias correction of different precipitation products in the Yangtze River basin[J]. Journal of China Hydrology，2023，43(1)：62-67.(in Chinese))

[13] WU H C，YANG Q L，LIU J M，et al. A spatiotemporal deep fusion model for merging satellite and gauge precipitation in China[J]. Journal of Hydrology，2020，584：124664.

[14] 张炜，沈明星，高成，等. 集合卡尔曼滤波与随机森林算法在异源遥感降水数据同化融合中的应用 [J]. 水电能源科学，2024，42(8)：11-16.(ZHANG W，SHEN M X，GAO C，et al. Application of ensemble Kalman filter and random forest algorithm in assimilation and fusion of heterogeneous remote sensing precipitation data[J]. Water Resources and Power，2024，42(8)：11-16.(in Chinese))

[15] CHEN C F，HE Q X，LI Y Y. Downscaling and merging multiple satellite precipitation products and gauge observations using random forest with the incorporation of spatial autocorrelation[J]. Journal of Hydrology，2024，632：130919.

[16] LE X H，KIM Y，van BINH D，et al. Improving rainfall-runoff modeling in the Mekong River basin using bias-corrected satellite precipitation products by convolutional neural networks[J]. Journal of Hydrology，2024，630：130762.

[17] 李伶杰，王银堂，唐国强，等. 考虑有雨无雨辨识的多源降水融合方法 [J]. 水科学进展，2022，33(5)：780-793.(LI L J，WANG Y T，TANG G Q，et al. An innovative multi- source precipitation merging method with the identification of rain and no rain[J]. Advances in Water Science，2022，33(5)：780-793.(in Chinese))

[18] 孙敏. 青藏高原腹地降水多源数据融合订正和时空变化研究 [D]. 南京：南京信息工程大学，2024.(SUN M. Study on fusion correction and temporal and spatial variation of precipitation multi-source data in the hinterland of Qinghai-Tibet Plateau[D]. Nanjing：Nanjing University of Information Science & Technology，2024.(in Chinese))

[19] WANG H，QIN H，LIU G J，et al. A novel feature attention mechanism for improving the accuracy and robustness of runoff forecasting[J]. Journal of Hydrology，2023，618：129200.

[20] 陈娟，徐琦，曹端祥，等. 基于多因子多模式集成的中长期径流预测模型 [J]. 水科学进展，2024，35(3)：408-419.(CHEN J，XU Q，CAO D X，et al. Medium and long-term runoff prediction model based on multi-factor and multi-model integration[J]. Advances in Water Science，2024，35(3)：408-419.(in Chinese))

[21] YAN Y H，LI X R，SUN W J，et al. Semi-surrogate modelling of droplets evaporation process via XGBoost integrated CFD simulations[J]. Science of the Total Environment，2023，895：164968.

[22] LUNDBERG S M，LEE S I. A unified approach to interpreting model predictions[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach，USA：ACM，2017：4768-4777.

[23] 周嘉月，卢麾，阳坤，等. 基于 CMIP6 的中高温升情景对中国未来径流的预估 [J]. 中国科学：地球科学，2023，53(3)：505-524.(ZHOU J Y，L H，Y K，et al. Projection of China's future runoff based on the CMIP6 mid-high warming scenarios[J]. SCIENTIA SINICA Terrae，2023，66(3)：528-546.(in Chinese))

[24] CAO C J，HE Y Y，CAI S Y. Probabilistic runoff forecasting considering stepwise decomposition framework and external factor integration structure[J]. Expert Systems with Applications，2024，236：121350.

[25] 黄志萍，任广成，夏军. 盛夏副高东西位置异常变化对我国气候的影响及预测研究 [J]. 海洋预报，2012，29(3)：53-61.（HUANG Z P，REN G C，XIA J. Impact of anomalous latitudinal position of western Pacific subtropical high on climate change in midsummer in China[J]. Marine Forecasts，2012，29(3)：53-61.（in Chinese））

[26] ZHANG Q，SHEN Z X，POKHREL Y，et al. Oceanic climate changes threaten the sustainability of Asia's water tower[J]. Nature，2023，615(7950)：87-93.

[27] 袁媛，丁婷，高辉，等. 我国南方盛夏气温主模态特征及其与海温异常的联系 [J]. 大气科学，2018，42(6)：1245-1262.（YUAN Y，DING T，GAO H，et al. Major modes of midsummer air temperature in Southern China and their relationship with sea surface temperature anomalies[J]. Chinese Journal of Atmospheric Sciences，2018，42(6)：1245-1262.（in Chinese））

[28] 崔乃文，范广洲. 极涡与南亚高压的关系及对我国降水的协同影响 [J]. 高原山地气象研究，2021，41(2)：1-9.（CUI N W，FAN G Z. The relationship between polar vortex and South Asian high and its synergistic influences on precipitation in China[J]. Plateau and Mountain Meteorology Research，2021，41(2)：1-9.（in Chinese））

[29] 晏红明，李刚，袁媛，等. 2021/2022 年东亚大陆冬季前暖后冷的环流差异及其成因 [J]. 地球物理学报，2023，66(10)：4026-4044.（YAN H M，LI G，YUAN Y，et al. Atmospheric circulation difference between warm in early winter and cold in late winter of 2021/2022 over East Asia continent and its causes[J]. Chinese Journal of Geophysics，2023，66(10)：4026-4044.（in Chinese））

Study on the bias correction method for monthly precipitation data considering teleconnection factors*

YAN Baowei[1,2]，CHANG Jianbo[1,2]，SUN Mingbo[1,2]，GU Donglin[1,2]，ZHOU Xuerui[1,2]，YANG Dongxu[1,2]

（1. *School of Civil and Hydraulic Engineering*，*Huazhong University of Science and Technology*，*Wuhan* 430074，*China*；2. *Hubei Key Laboratory of Digital River Basin Science and Technology*，*Huazhong University of Science and Technology*，*Wuhan* 430074，*China*）

Abstract：The correction of multi-source precipitation data is critical for analyzing and modeling hydrological patterns in data-scarce regions. However，existing correction methods often overlook the influence of climatic factors. To address this，a precipitation bias correction model was developed based on BiLSTM，integrating multi-source precipitation datasets (ERA5，ERA5-Land，MSWEP-V2，and PERSIANN-CDR) with teleconnection factors. The XGBoost-SHAP model was employed for feature selection and causality analysis，while a Bayesian Optimization (BO) strategy was applied to identify the optimal combination of model hyperparameters to further enhance correction accuracy. Using the Upper Hanjiang River basin as the study area，multi-source precipitation data bias correction was conducted. The results indicate: ①Atmospheric circulation factors are the primary influences on precipitation formation in the Upper Hanjiang River basin，with the position index of the Northern Hemisphere Subtropical High Ridge having the most significant impact. ②Compared with traditional statistical methods，the BO-BiLSTM model，while slightly less effective than the optimal parameter transformation method，provides greater flexibility in incorporating multiple influencing factors. ③After considering teleconnection factors，the Nash-Sutcliffe efficiency coefficient of the corrected multi-source precipitation data during the test period improved by an average of 5.4%，the mean squared error decreased by 24.6% on average，and the Kling-Gupta efficiency coefficient increased by 10.5% on average. These findings offer a practical technical solution for high-precision monthly precipitation estimation and extension in data-scarce regions.

Key words：monthly precipitation；bias correction；teleconnection factors；multi-source precipitation dataset；BiLSTM model

* The study is financially supported by the National Key R&D Program of China（No. 2021YFC3200301）and the National Natural Science Foundation of China（No. 52079054）.