曹子恒, 李永坤, 胡义明, 等. 基于机器学习模型的数值降雨预报校正[J]. 南水北调与水利科技(中英文), 2023, 21(5): 843-861, 950.
CAO Z H, LI Y K, HU Y M, et al. Numerical rainfall forecast correction based on machine learning model[J]. South-to-North Water Transfers and Water Science & Technology, 2023, 21(5): 843-861, 950. (in Chinese)

基于机器学习模型的数值降雨预报校正

曹子恒[1], 李永坤[2], 胡义明[1], 卢亚静[2], 温骐宇[1], 杨晨曦[1], 陈钰[1], 郭举坤[1]

（1. 河海大学水文水资源学院, 南京 210098; 2. 北京市水科学技术研究院, 北京 100000）

摘要: 以潮白河流域 12 个站点为研究对象, 选取 12 个站点的未来 12 h 不同预见期的预报降水数据, 构建基于支持向量机 (support vector machine, SVM) 模型、随机森林 (random forest, RF) 模型和多层感知机 (multilayer perceptron, MLP) 模型的不同预见期预报降雨校正模型, 模型输入为站点对应网格及其周边 8 个网格的降雨预报数据, 模型参数采用贝叶斯优化技术进行估计。利用均方根误差和确定性系数评估各模型对不同预见期预报降水的校正效果。结果表明: 未经校正的原始预报在不同预见期的预报精度均较差; 各个误差校正模型在率定期与验证期对不同预见期降雨均具有较好的校正效果; 经 SVM、RF 和 MLP 模型校正后, 均方根误差的平均值在率定期分别降低了: 54.2%、50.0% 和 20.8%, 在验证期分别降低 42.9%、33.3% 和 14.3%; 确定性系数的平均值在率定期与验证期也均有显著提高; 3 个误差校正模型中, SVM 模型表现最优, RF 模型次之。研究成果可为其他流域数值降雨预报数据校正提供参考。

关键词: 预报降雨校正; 支持向量机; 随机森林; 多层感知机; 潮白河流域

中图分类号: TV213 **文献标志码:** A **DOI:** 10.13476/j.cnki.nsbdqk.2023.0083

降雨是形成洪水的直接因素, 精准的长预见期降雨预报数据与水文模型结合, 是提高洪水预报准确性和增长预见期的关键, 可为防洪减灾争取更长的应急响应时间[1-5]。现代降雨预报数据主要来源于气象雷达、卫星云图和数值天气预报产品等。尽管过去几十年里, 气象观测技术和设备有了长足进步, 但由于大气系统的混沌非线性、大气初始资料误差以及模式误差的存在, 降雨预报产品不可避免地具有较大的误差与局限性, 需要经过有效校正, 以提高其精确性和可靠性[6-9]。

近些年来, 随着大数据挖掘技术、机器学习算法及计算机软硬件环境的快速发展, 采用机器学习模型进行降雨预报校正成为一种有效途径, 受到越来越多的关注[10-16]。如: 疏杏胜等[10] 利用人工神经网络、极限学习机及支持向量机 (support vector machine, SVM) 模型对桓仁水库流域未来 1~3 d 降雨进行多模式集成预报, 结果显示多模式集成的效果要优于单一模型, 且 SVM 模型对降雨预报精度改善最为明显; Sun 等[12] 使用随机森林 (random forest, RF) 模型对 1951—2020 年青藏高原上游 11 个流域的 ERA5 降水数据进行校正并构建了网格化的日降水数据集, 结果显示校正后的降水数据与观测结果吻合较好; Appiah 等[14] 应用机器学习算法对加纳不同地区的降雨量进行了预测, 结果表明 RF 和多层感知机 (multilayer perceptron, MLP) 表现良好, 而 k 近邻算法表现较差。

本文以潮白河流域 12 个站点未来 12 h 降雨预报校正为研究对象, 构建基于支持向量机、随机森林和多层感知机算法的 3 种校正模型, 对不同预见期的降雨预报数据进行校正, 并评估各模型的校正效果。

1 模型方法

1.1 模型原理

1.1.1 随机森林

随机森林 (RF) 回归是一种引导聚合算法, 其使

收稿日期: 2023-06-08 **修回日期:** 2023-08-15 **网络出版时间:** 2023-09-28
网络出版地址: https://link.cnki.net/urlid/13.1430.TV.20230927.1651.004
基金项目: 国家自然科学基金重点项目 (41730750); 中央高校基本科研业务费专项项目 (B220202031)
作者简介: 曹子恒(1998—), 男, 江苏泰州人, 主要从事水文水资源研究。E-mail: 221301030002@hhu.edu.cn
通信作者: 胡义明(1986—), 男, 江苏宿迁人, 副教授, 博士, 主要从事水文水资源研究。E-mail: yiming.hu@hhu.edu.cn

用决策树作为基础学习器[17-18]。在对数据集进行自助抽样基础上,结合决策树算法,创建不同的回归预测模型,并通过综合集成不同模型以获得最终预测结果。RF 算法包含以下主要步骤:利用自助抽样 (bootstrap) 技术从原始数据集中抽样生成多个子数据集;针对每个子数据集,使用随机特征选择方法筛选特征变量,用于训练决策树模型;对于每棵决策树,采用 CART 算法进行构建,选择最佳的特征作为分裂节点,并在每个叶节点上预测该节点对应样本的输出值;对每棵决策树的预测结果进行综合以获得最终预测结果。随机森林模型的整体结构见图 1。

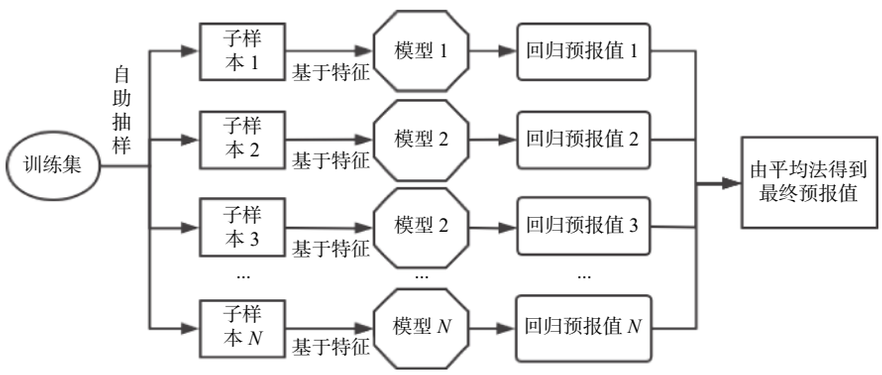


图 1　随机森林回归模型结构

Fig. 1　Structure of random forest regression model

1.1.2　多层感知机

多层感知机 (MLP) 是一种前馈神经网络,它由输入层、隐藏层和输出层多个神经元层组成,隐藏层可以是多层[19-20]。每个神经元都与前一层和后一层的所有神经元相连,其结构见图 2。

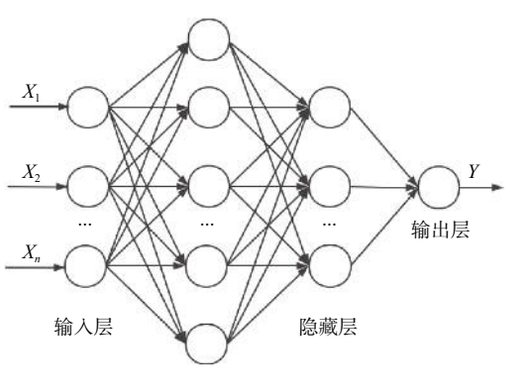


图 2　多层感知机模型结构

Fig. 2　Structure of multilayer perceptron model

应用 MLP 模型进行回归问题分析时,输出层通常是一个线性函数,用于预测目标变量的数值。每个神经元都有一组权重,用于将输入值映射到神经元的输出。每个神经元还有一个偏置项,通过激活函数(如 Sigmod、Tanh 和 ReLU)进行激活。模型使用损失函数来度量预测值与真实值之间的误差,常用的损失函数包括均方误差和平均绝对误差(MAE)等。采用 Adam 优化器和随机梯度下降算法等方法优化模型参数以最小化损失函数,并通过反向传播算法更新每个神经元的权重和偏置,直到达到一定的训练准确率或最小化损失函数为止。MLP 模型的一般表达形式为

$$\boldsymbol{Y} = f(w_i\boldsymbol{X} + \boldsymbol{B}_i) \tag{1}$$

式中:$\boldsymbol{Y}$为输出值;f为激活函数;w_i为全连接层的权重;$\boldsymbol{X}$为输入层变量矩阵;$\boldsymbol{B}_i$为偏置项。

1.1.3　支持向量机

支持向量机 (SVM) 回归的基本思想是寻找一个线性或非线性超平面,使得这个超平面与数据点之间的间隔最大化,并找到一组支持向量,使它们距离超平面最近且落在间隔边界上[21-24]。在 SVM 中,为了使预测值与真实值之间误差最小且模型具有较低复杂度,通常会选择合适的核函数将数据从输入空间映射到高维特征空间,在特征空间中求解最优超平面。在线性回归模型中,SVM 使用线性核函数,在非线性回归中,通常使用径向基函数等核函数。SVM 回归模型的一般表达形式为

$$\min: \frac{1}{2}\|\omega\|^2 \tag{2}$$

$$\text{s.t.} \begin{cases} y_i - \omega x - b \leqslant \varepsilon \\ \omega x + b - y_i \leqslant \varepsilon \end{cases} \tag{3}$$

式中:ω是权重向量;b是偏差项;y_i是第 i 个样本的真实输出值;ε是间隔大小;x是需要求解的一组最优向量。目标是最小化权重向量的平方范数,同时保证每个样本的预测误差不超过ε。

1.2　模型参数估计

采用贝叶斯优化技术估计上述各模型的参数。

相比于传统超参数优化方法，贝叶斯优化技术考虑了各个超参数的先验分布，通过不断更新各参数先验分布进而获得超参数的最佳估计[25-26]。贝叶斯优化技术在更高可能性的区域内进行参数采样，避免遍历整个超参数空间，具有更好的估计效率和准确率。此外，贝叶斯优化技术可以对超参数空间进行自适应搜索，更容易找到全局最优解。在设置各模型中关键超参数的取值范围基础上，采用贝叶斯优化技术可进行参数的高效估计。表 1 给出了随机森林模型中关键超参数的取值范围设定。

表 1　随机森林模型中关键超参数取值范围

Tab. 1　Value range of some hyperparameters in the RF

n_estimators	max_features	max_depth	min_samples_split	min_samples_leaf
[10~100]	['sqrt', 'log2', auto']	[2~20]	[2~20]	[1~10]

表 1 中：n_estimators 为随机森林中树的数量；max_features 用于确定每棵决策树节点上用于拆分的特征数量范围或具体取值；max_depth 为决策树的最大深度；min_samples_split 为决策树节点分裂所需的最小样本数；min_samples_leaf 为叶子节点所需的最小样本数，当一个节点的样本数小于 min_samples_leaf 时，该节点会被剪枝成为叶子节点。

表 2 给出了多层感知机模型中关键超参数的取值范围设定。

表 2　多层感知机模型中关键超参数取值范围

Tab. 2　Value range of some hyperparameters in the MLP

hidden_layer_sizes	activation	alpha	learning_rate	solver	max_iter
[10~500]	['relu', 'tanh', 'logistic']	[0.000 001~100]	['constant', 'adaptive']	['adam']	[100~5 000]

考虑到本研究中数据集的规模较小，采用单层隐藏层的 MLP 模型，减少模型复杂度和过拟合风险。hidden_layer_sizes 代表隐藏层中神经元的数量；activation 为激活函数；alpha 为 L2 正则化系数；learning_rate 为学习率；solver 为优化器类型；max_iter 为最大迭代次数。

表 3 给出了支持向量机模型中关键超参数的取值范围设定。

表 3　支持向量机模型中关键超参数取值范围

Tab. 3　Value range of some hyperparameters in the SVM

C	gamma	kernel	epsilon
Real(1e−6, 1e+4, 'log-uniform')	Real(1e−6, 1e+4, 'log-uniform')	['rbf']	Real(0, 1, 'uniform')

表 3 中：C 为正则化参数，用于控制分类器对误分类样本的惩罚程度；gamma 为核函数参数，控制样本点映射到高维空间后的分布特征，Real(1e−6, 1e+4, 'log-uniform') 定义了一个范围从 1e−6 到 1e+4 连续的实数型搜索空间，并在该范围内使用对数均匀分布方式进行搜索；kernel 为核函数类型，本次选用径向基函数；epsilon 为浮点数，作用于损失函数。

1.3　模型评价指标

采用均方根误差 E_{RMS} 和确定性系数 R^2 评估原始降雨预报精度及各个模型对不同预见期预报降雨的校正效果。

均方根误差的计算公式为

$$E_{\mathrm{RMS}}=\sqrt{\frac{\sum(y_{\mathrm{pred}}-y_{\mathrm{true}})^2}{N}} \tag{4}$$

式中：E_{RMS}为均方根误差；N为样本数；y_{pred}为模型预测值；y_{true}为实测值。E_{RMS} 值越小，表示模型的预报结果越接近实测，模型精度越高。

确定性系数的计算公式为

$$R^2=1-\frac{\sum(y_{\mathrm{pred}}-y_{\mathrm{true}})^2}{\sum(y_{\mathrm{true}}-\bar{y}_{\mathrm{true}})^2} \tag{5}$$

式中：R^2为确定性系数；y_{pred}为模型预测值；y_{true}为实测值；$\bar{y}_{\mathrm{true}}$为实测值均值。R^2 越接近 1，表明模型的预测精度越高。

2　结果分析

2.1　数据资料信息

以潮白河流域（北京市内区域）12 个站点（图 3）

未来 12 个不同预见期的预报降雨校正为研究对象，采用的数据为 2021 年和 2022 年汛期 6—9 月的逐小时数据，在每个整点数值预报产品都会滚动预报未来 12 h 逐小时的降雨，空间分辨率为 500 m×500 m。

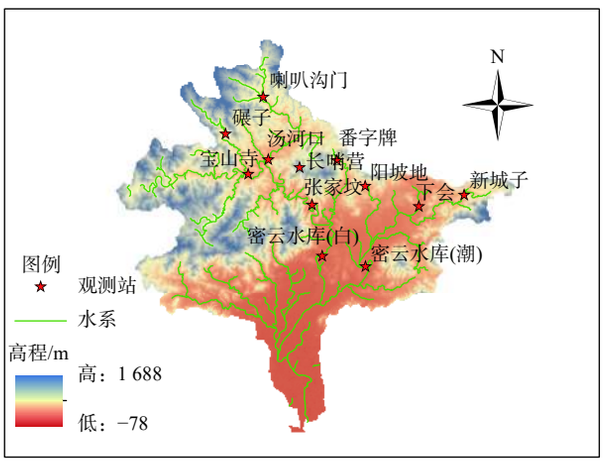


图 3　研究区观测站点

Fig. 3　Overview of the study area

2.2　校正效果分析

校正模型是分预见期构建的，即每个站点 12 个预见期的降雨校正需要构建 12 个不同的校正模型。模型的输入为站点对应的网格和其周边 8 个网格共 9 个网格的预报降雨数据，即将 9 个网格的预报降雨数据作为模型的输入。9 个网格预报降雨的平均值作为该站点对应的原始预报降雨值。采用 80% 的样本数据进行模型率定，剩余 20% 的样本数据进行模型验证。考虑到不同时刻降雨量的数值在量级上差别较大，采用下式对降雨数据进行标准化处理。

$$Z=\frac{x-\mu}{\sigma} \tag{6}$$

式中：Z为标准化后的变量；x为输入变量；μ为输入变量的均值；σ为输入变量的标准差。

图 4 和图 5 分别给出了率定期各站点在不同预见期下原始预报降雨及经各模型校正后预报降雨的均方根误差和确定性系数，可以看出，经机器学习模型校正后，各站不同预见期预报降雨数据的均方根误差值均有所减小，而确定性系数值均有所提升，即各个模型对不同预见期降雨均具有较好的校正效果，可提高降雨预报精度，其中，支持向量机对数据的校正效果最好。

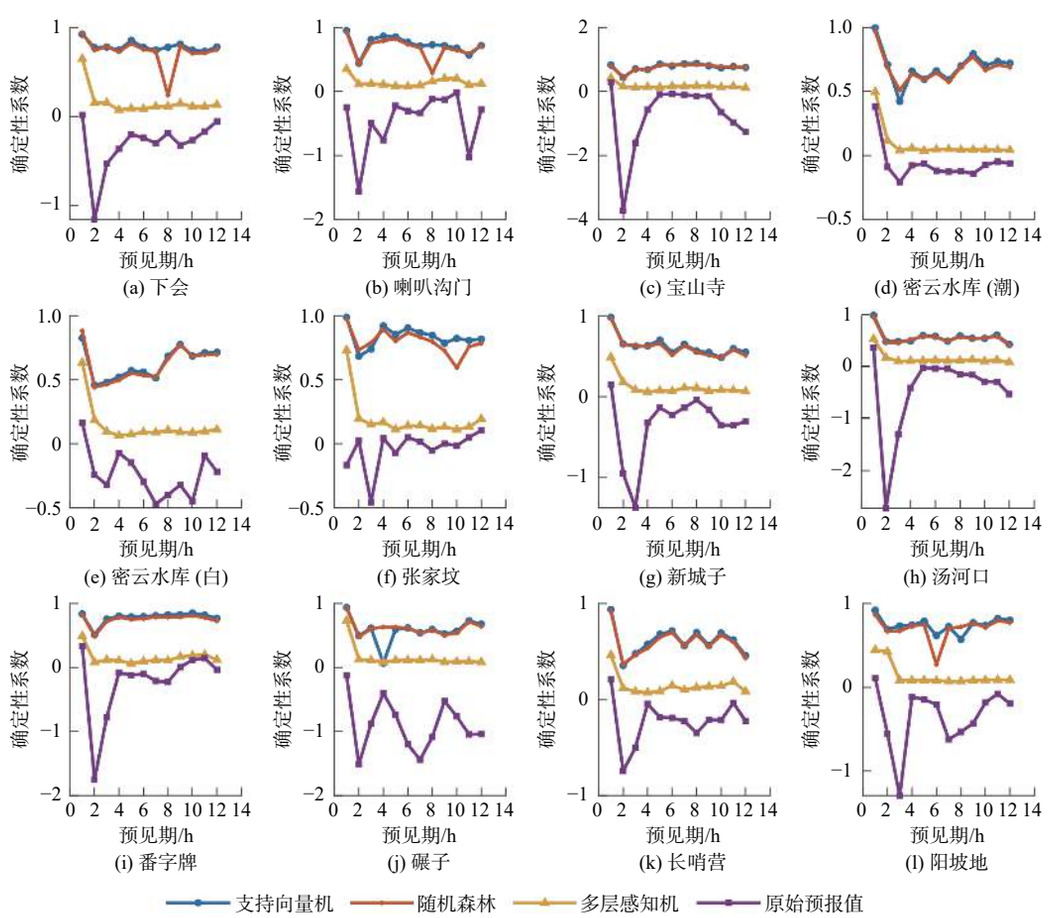


图 4　率定期各站点确定性系数

Fig. 4　R^2 at each station in train period

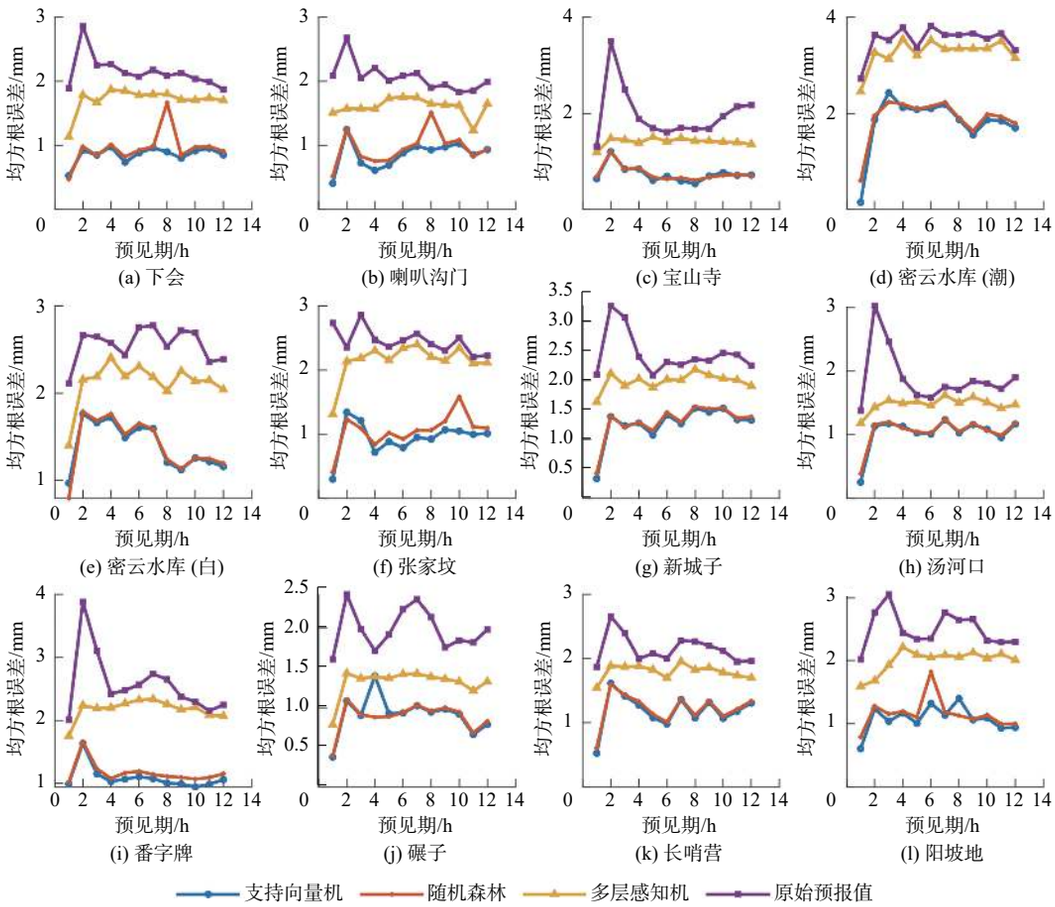


图 5　率定期各站点均方根误差

Fig. 5　E_{RMS} at each station in train period

表 4 给出了率定期不同预见期下 12 个站点对应的均方根误差的均值和确定性系数的均值,可以看出:经过 3 个模型校正后,每个预见期的确定性系数都有了明显提升,均方根误差都降低显著;3 个模型中,支持向量机模型表现最优,随机森林模型次之。

表 4　率定期 12 个站点均方根误差和确定性系数的均值

Tab. 4　The mean value of E_{RMS} and R^2 of 12 stations in train period

预见期/h	R^2				E_{RMS}/mm			
	原始值	支持向量机	随机森林	多层感知机	原始值	支持向量机	随机森林	多层感知机
1	0.12	0.88	0.79	0.46	1.36	0.52	0.67	1.11
2	−1.29	0.56	0.51	0.16	2.32	1.08	1.18	1.60
3	−0.51	0.59	0.54	0.09	2.55	1.34	1.41	2.06
4	−0.14	0.63	0.58	0.07	1.82	0.99	1.10	1.65
5	−0.12	0.61	0.46	0.07	2.13	1.23	1.48	1.99
6	−0.37	0.56	0.40	0.06	2.09	1.21	1.43	1.77
7	−0.57	0.58	0.37	0.04	1.91	1.03	1.23	1.58
8	−0.34	0.48	0.28	0.05	2.17	1.33	1.61	1.87
9	−0.37	0.51	0.38	0.05	2.07	1.26	1.42	1.78
10	−0.36	0.54	0.40	0.06	2.14	1.25	1.42	1.85
11	−0.23	0.45	0.33	0.03	2.42	1.60	1.81	2.19
12	−0.13	0.44	0.36	0.11	2.40	1.71	1.83	2.16

图 6 和图 7 分别给出了验证期各站点在不同预见期下原始预报降雨及经各模型校正后预报降雨的均方根误差和确定性系数,可以看出:各指标反映的结论与率定期一致,即经各模型校正后不同预见期降雨的精度均有所提高;3 个模型中,支持向量机模型表现最优。

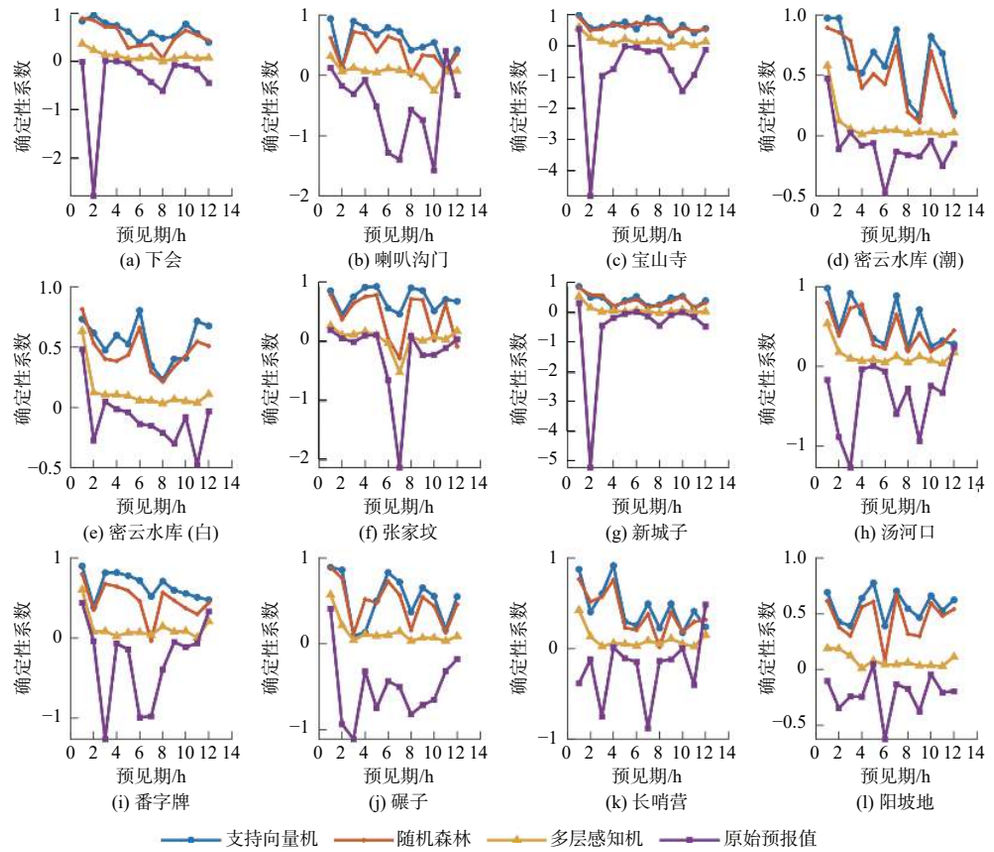


图 6　验证期各站点确定性系数

Fig. 6　R^2 at each station in test period

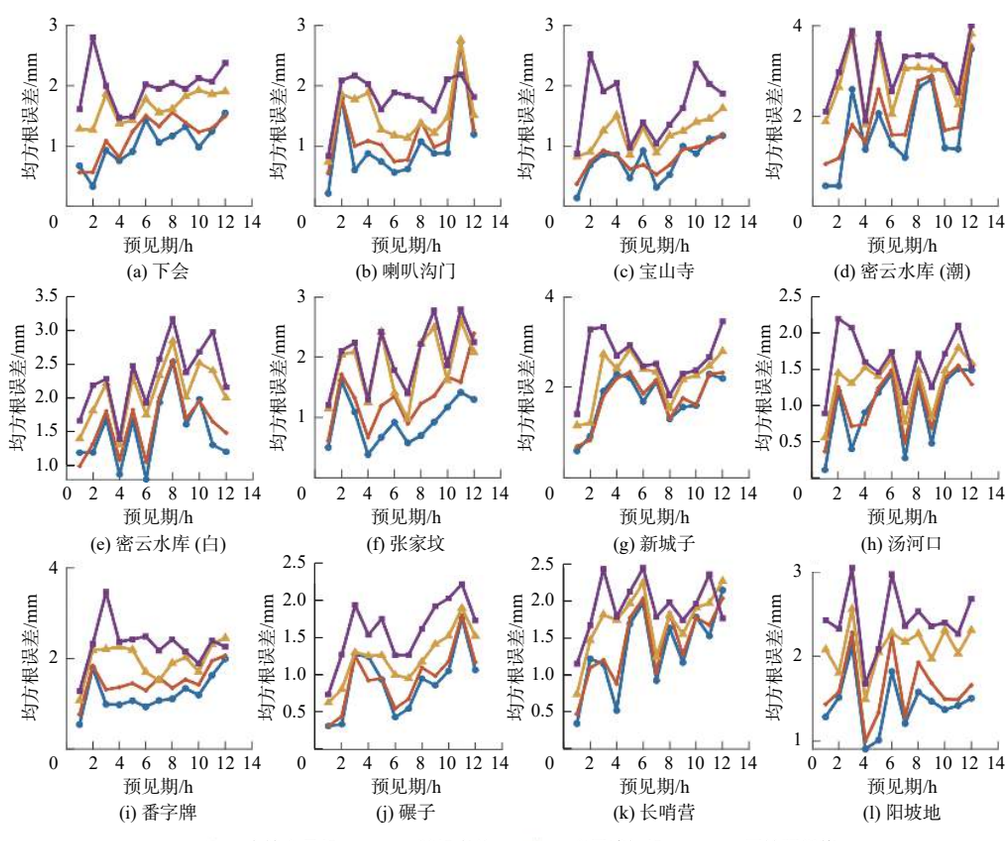


图 7　验证期各站点均方根误差

Fig. 7　E_{RMS} at each station in test period

表 5 给出了验证期不同预见期下 12 个站点对应的均方根误差的均值和确定性系数的均值，可以看出：与率定期结论一致，经过 3 个模型校正后，每个预见期的确定性系数都有明显提升，均方根误差都降低显著；3 个模型中，支持向量机模型最优，随机森林模型次之。

表 5　验证期 12 个站点均方根误差和确定性系数的均值

Tab. 5　The mean value of E_{RMS} and R^2 of 12 stations in test period

预见期/h	R^2				E_{RMS}/mm			
	原始值	支持向量机	随机森林	多层感知机	原始值	支持向量机	随机森林	多层感知机
1	0.12	0.88	0.79	0.46	1.36	0.52	0.67	1.11
2	−1.29	0.56	0.51	0.16	2.32	1.08	1.18	1.60
3	−0.51	0.59	0.54	0.09	2.55	1.34	1.41	2.06
4	−0.14	0.63	0.58	0.07	1.82	0.99	1.10	1.65
5	−0.12	0.61	0.46	0.07	2.13	1.23	1.48	1.99
6	−0.37	0.56	0.40	0.06	2.09	1.21	1.43	1.77
7	−0.57	0.58	0.37	0.04	1.91	1.03	1.23	1.58
8	−0.34	0.48	0.28	0.05	2.17	1.33	1.61	1.87
9	−0.37	0.51	0.38	0.05	2.07	1.26	1.42	1.78
10	−0.36	0.54	0.40	0.06	2.14	1.25	1.42	1.85
11	−0.23	0.45	0.33	0.03	2.42	1.60	1.81	2.19
12	−0.13	0.44	0.36	0.11	2.40	1.71	1.83	2.16

以密云水库 (潮) 第 1 个预见期的预报降雨为例，绘制模型校正前后预报降雨与实测降雨的散点图见图 8，可以看出：未经校正的原始预报降雨与实测降雨间的点据较为散乱，经模型校正后，预报降雨与实测降雨的点据趋向于聚集在 1∶1 线附近，表明校正后的预报降雨精度有所提高，更接近实测降雨；无论是率定期还是验证期，支持向量机模型效果最优，随机森林模型次之。

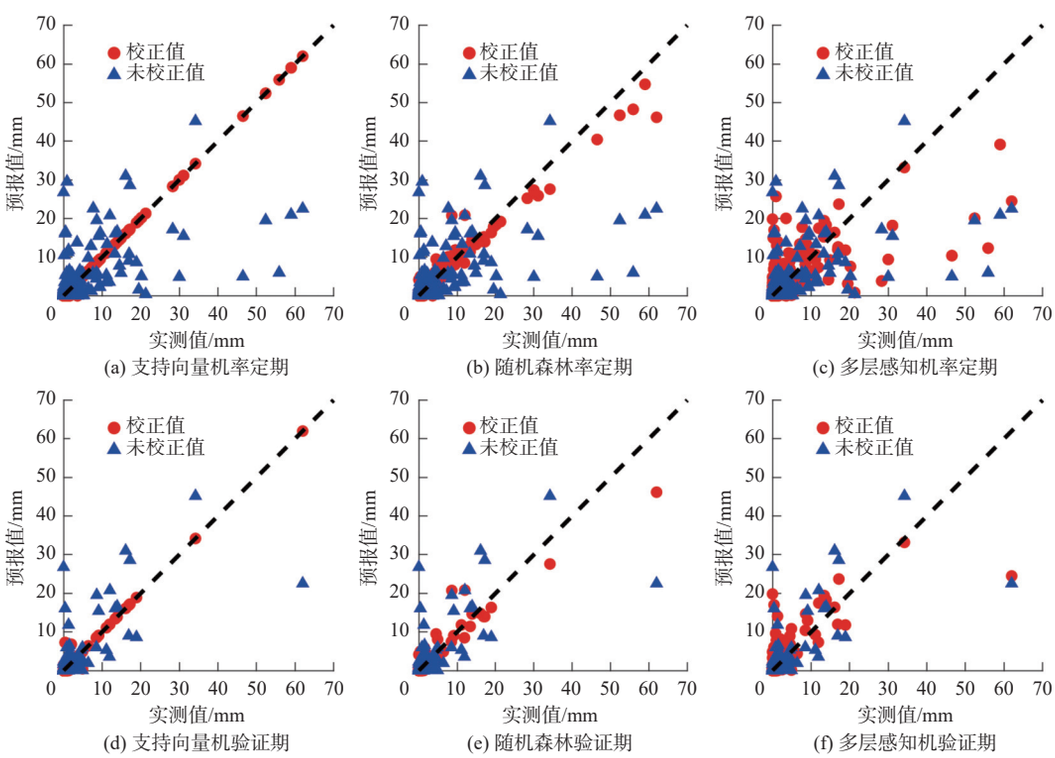


图 8　模型校正前后预报降雨与实测降雨的散点图

Fig. 8　Scattered plot of forecast rainfall and measured rainfall before and after model correction

SVM 模型优于其他两个模型的主要原因[27-29]：SVM 采用结构风险最小化原则进行建模，在最小化训练误差的同时减小泛化误差，有效地避免过拟合现象，使得 SVM 具有较强的泛化能力；SVM 算法通过最大化间隔来确定决策边界，对离群点具有较好的鲁棒性，可以有效避免异常点对结果的影响；SVM 的性能受训练样本数目和样本分布特征的综合影响，在样本数较小情形下，SVM 能更好地处理数据分布不均情况，更适用于小样本数据建模。

3 结 论

本文基于 SVM、RF 和 MLP 模型，结合贝叶斯优化技术，构建不同预见期预报降雨数据的校正模型，对潮白河流域 12 个站点未来 12 个不同预见期的预报降雨数据进行校正分析。

采用 E_{RMS} 和 R^2评估原始预报、经 SVM、RF 和 MLP 模型校正后预报的效果，就 12 个预见期对应 R^2的平均值而言，在率定期分别为−0.37、0.69、0.67 和 0.15，在验证期分别为−0.36、0.57、0.45 和 0.11，各校正模型对各站不同预见期的预报降雨均具有较好的校正效果；就 E_{RMS} 指标而言，经 SVM、RF 和 MLP 模型校正后，E_{RMS} 平均值在率定期分别降低 54.2%、50.0% 和 20.8%，在验证期分别降低 42.9%、33.3% 和 14.3%。3 个校正模型中，SVM 模型最优，RF 模型次之。

相较于机器学习中常用的网格搜索、随机搜索等参数优化方法，本次采用的贝叶斯优化方法能够在相对较少的运行负荷下获得参数最优解，且可获得参数的概率分布估计，以分析参数估计的不确定性。此外，基于参数的概率分布估计，结合马尔科夫链蒙特卡洛抽样技术，可获得多套模型参数集，进而可用于分析降雨校正的不确定性。后续工作将对此进一步深入研究。

参考文献：

[1] 黄一昕, 王钦钊, 梁忠民, 等. 洪水预报实时校正技术研究进展[J]. 南水北调与水利科技 (中英文), 2021, 19(1): 12-35. DOI: 10.13476/j.cnki.nsbdqk.2021.0002.

[2] 韦经豪, 黄迎春, 姚成. 降水预报产品在不同水文气象分区中小流域的适应性评估[J]. 南水北调与水利科技 (中英文), 2022, 20(6): 1208-1219. DOI: 10.13476/j.cnki.nsbdqk.2022.0119.

[3] 邸苏闯, 李卓蔓, 刘玉, 等. 基于气象雷达反演和云图外推法的临近期降雨预报方法研究[J]. 水利水电技术 (中英文), 2022, 53(5): 13-21. DOI: 10.13928/j.cnki.wrahe.2022.05.002.

[4] 刘志雨, 刘玉环, 孔祥意. 中小河流洪水预报预警问题与对策及关键技术应用[J]. 河海大学学报 (自然科学版), 2021, 49(1): 1-6. DOI: 10.3876/j.issn.1000-1980.2021.01.001.

[5] 张玉兰, 张卫国, 贾本有, 等. 基于防汛需求的降雨预报精度评估方法[J]. 南水北调与水利科技 (中英文), 2021, 19(2): 293-300. DOI: 10.13476/j.cnki.nsbdqk.2021.0031.

[6] 胡义明, 梁忠民, 蒋晓蕾, 等. GFS集合降雨预报的校正后处理研究[J]. 南水北调与水利科技(中英文), 2019, 17(1): 15-19. DOI: 10.13476/j.cnki.nsbdqk.2019.0003.

[7] 唐榕, 王运涛, 李敏, 等. ECMWF降雨预报信息不同利用形式精度评估[J]. 中国农村水利水电, 2020, 453(7): 1-5. DOI: 10.3969/j.issn.1007-2284.2020.07.001.

[8] 温立成, 李致家. 校正后的降雨格点预报在洪水预报中的应用[J]. 水电能源科学, 2010, 28(4): 1-4. DOI: 10.3969/j.issn.1000-7709.2010.04.001.

[9] 吴旭树, 王兆礼, 陈柯兵, 等. 基于大气环流和海温场的降水组合预报模型[J]. 水资源保护, 2022, 38(6): 81-87. DOI: 10.3880/j.issn.1004-6933.2022.06.011.

[10] 疏杏胜, 王子茹, 李福威, 等. 基于机器学习模型的短期降雨多模式集成预报[J]. 南水北调与水利科技 (中英文), 2020, 18(1): 42-50. DOI: 10.13476/j.cnki.nsbdqk.2020.0006.

[11] ORTIZ-GARCIA E G, SALCEDO-SANZ S, CASANOVA-MATEO C. Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data[J]. Atmospheric Research, 2014, 139: 128-136. DOI: 10.1016/j.atmosres.2014.01.012.

[12] SUN H, YAO T D, SU F G, et al. Corrected ERA5 precipitation by machine learning significantly improved flow simulations for the Third Pole basins[J]. Journal of Hydrometeorology, 2022, 23(10): 1663-1679. DOI: 10.1175/JHM-D-22-0015.1.

[13] ADARYANI F R, MOUSAVI S J, JAFARI F. Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN[J]. Journal of Hydrology, 2022, 614: 128463. DOI: 10.1016/j.jhydrol.2022.128463.

[14] APPIAH-BADU N K A, MISSAH Y M, AMEKUDZI L K, et al. Rainfall prediction using machine learning algorithms for the various ecological zones of Ghana[J]. IEEE Access, 2021, 10: 5069-5082. DOI: 10.1109/ACCESS.2021.3139312.

[15] KO C M, JEONG Y Y, LEE Y M, et al. The development of a quantitative precipitation forecast correction technique based on machine learning for hydrological applications[J]. Atmosphere, 2020, 11(1): 111. DOI: 10.3390/atmos11010111.

[16] LI H Y, ZHANG Y, LEI H J, et al. Machine learning-based bias correction of precipitation measurements at high altitude[J]. Remote Sensing, 2023, 15(8): 2180. DOI: 10.3390/rs15082180.

[17] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45: 5-32. DOI: 10.1023/A:1010933404324.

[18] ZOUNEMAT-KERMANI M, BATELAAN O, FADAEE M, et al. Ensemble machine learning paradigms in hydrology: A review[J]. Journal of Hydrology, 2021, 598: 126266. DOI: 10.1016/j.jhydrol.2021.126266.

[19] CHOUBIN B, KHALIGHI-SIGAROODI S, MALEKIAN A, et al. Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals[J]. Hydrological Sciences Journal, 2016, 61(6): 1001-1009. DOI: 10.1080/02626667.2014.966721.

[20] UYSAL G. Product-and hydro-validation of satellite-based precipitation data sets for a poorly gauged snow-fed basin in Turkey[J]. Water, 2022, 14(17): 2758. DOI: 10.3390/w14172758.

[21] 胡义明, 陈腾, 罗序义, 等. 基于机器学习模型的淮河流域中长期径流预报研究[J]. 地学前缘, 2022, 29(3): 284-291. DOI: 10.13745/j.esf.sf.2021.10.2.

[22] 牛欣怡, 鲁程鹏, 卢佳赟, 等. 机器学习模型在地下水埋深模拟中的适应性分析[J]. 河海大学学报 (自然科学版), 2022, 50(4): 74-82. DOI: 10.3876/j.issn.1000-1980.2022.04.010.

[23] 张岩, 杨明祥, 雷晓辉, 等. 基于PCA-PSO-SVR的丹江口水库年径流预报研究[J]. 南水北调与水利科技, 2018, 16(5): 35-40. DOI: 10.13476/j.cnki.nsbdqk.2018.0122.

[24] YIN G, YOSHIKANE T, YAMAMOTO K, et al. A support vector machine-based method for improving real-time hourly precipitation forecast in Japan[J]. Journal of Hydrology, 2022, 612: 128125. DOI: 10.1016/j.jhydrol.2022.128125.

[25] 占敏, 薛惠锋, 王海宁, 等. 贝叶斯神经网络在城市短期用水预测中的应用[J]. 南水北调与水利科技, 2017, 15(3): 73-79. DOI: 10.13476/j.cnki.nsbdqk.2017.03.013.

[26] MONEGO V S, ANOCHI J A, DE CAMPOS VELHO H F. South America seasonal precipitation prediction by gradient-boosting machine-learning approach[J]. Atmosphere, 2022, 13(2): 243. DOI: 10.3390/atmos13020243.

[27] KARAMIZADEH S, ABDULLAH S M, HALIMI M, et al. Advantage and drawback of support vector machine functionality[C]//2014 international conference on computer, communications, and control technology (I4CT). IEEE, 2014: 63-65. DOI: 10.1109/I4CT.2014.6914146. DOI: 10.1109/I4CT.2014.6914146

[28] ANANDHI A, SRINIVAS V V, NANJUNDIAH R S, et al. Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine[J]. International Journal of Climatology:A Journal of the Royal Meteorological Society, 2008, 28(3): 401-420. DOI: 10.1002/joc.1529.

[29] LI G, CHANG W, YANG H. A novel combined prediction model for monthly mean precipitation with error correction strategy[J]. IEEE Access, 2020, 8: 141432-141445. DOI: 10.1109/ACCESS.2020.3013354.

• 译文 •

Numerical rainfall forecast correction based on machine learning model

CAO Ziheng[1], LI Yongkun[2], HU Yiming[1], LU Yajing[2],
WEN Qiyu[1], YANG Chenxi[1], CHEN Yu[1], GUO Jukun[1]

（1. *College of Hydrology and Water Resources*, *Hohai University*, *Nanjing* 210098, *China*;
2. *Beijing Water Science & Technology Institute*, *Beijing* 100000, *China*）

Abstract: Rainfall is a direct factor in the formation of flood, and the combination of accurate rainfall forecast data in the long forecast period and hydrological model is the key to improve the accuracy of flood forecast and increase the forecast period, which can strive for a longer emergency response time for flood control and disaster reduction. Rainfall forecast data mainly come from meteorological radar, satellite cloud image and numerical weather forecast products. Although the meteorological observation technology and equipment have made great progress in the past few decades, due to the chaos of atmospheric system, the error of atmospheric initial data and the error of model, the rainfall forecast products inevitably have large errors and limitations, and need to be effectively corrected to improve its accuracy and reliability. The research took 12 stations in Chaobai River basin as the research object, the forecast precipitation data of 12 stations in different forecast periods in the next 12 hours were selected. Rainfall error correction models based on support vector machine, random forest and multilayer perceptron in different forecast periods were constructed. The model input is the rainfall forecast data of the corresponding grid of the station and its surrounding 8 grids, and the model parameters are estimated by Bayesian optimization technology. The root mean square error and deterministic coefficient indexes were used to evaluate the correction effect of each model on precipitation forecast in different forecast periods. The results showed that the prediction accuracy of uncorrected original forecast was poor in different forecast periods. Each error correction model has a good correction effect on rainfall in different forecast periods. After correction by support vector machine model, random forest model and multilayer perceptron model, the average root mean square error decreases by 54.2%, 50.0% and 20.8%, respectively. During the validation period, the reduction was 42.9%, 33.3% and 14.3%, respectively. The average certainty coefficient also increased significantly in both the rate period and the validation period. Among the three error correction models, support vector machine model is the best, followed by random forest model. Based on support vector machine, random forest and multi-layer perceptron model, combined with Bayesian optimization technology, the error correction models of forecast rainfall data in different forecast periods were constructed to correct and analyze the forecast rainfall data of 12 stations in the Chaobai River basin in 12 different forecast periods. The root mean square error and deterministic coefficient were used. The correction effect is good and the accuracy of rainfall forecast is improved, and it can be used as a reference for the numerical rainfall correction of other watershed stations.

Key words: rainfall forecast correction; support vector machine; random forest; multilayer perceptron; Chaobai River basin

Chinese Library Classification No. : TV213 **Document Code:** A

Rainfall is a direct factor in the formation of flood, and the combination of accurate rainfall forecast data in the long forecast period and hydrological model is the key to improve the accuracy of flood forecast and increase the forecast period, which can strive for a longer emergency response time for flood control and disaster reduction [1-5]. Modern rainfall forecast data mainly come from meteorological radar, satellite cloud image and numerical weather forecast products. Although the meteorological observation technology and equipment have made great progress in the past few decades, due to the chaotic nonlinearity of

Received: 2023-06-08 **Revised:** 2023-08-15
Fund: National Natural Science Foundation of China (41730750); Central University Basic Research Business Fee Special Project (B220202031)
Author's brief: CAO Ziheng (1998–), male, born in Taizhou, Jiangsu Province, mainly engaged in hydrology and water resources research. E-mail: 221301030002@hhu.edu.cn
Corresponding author: HU Yiming (1986–), male, born in Suqian, Jiangsu Province, associate professor, doctor, mainly engaged in hydrology and water resources research. E-mail: yiming.hu@hhu.edu.cn

atmospheric system, the error of atmospheric initial data and the error of model, the rainfall forecast products inevitably have large errors and limitations, and need to be effectively corrected to improve its accuracy and reliability [6-9].

In recent years, with the rapid development of big data mining technology, machine learning algorithms and computer software and hardware environments, the use of machine learning models for rainfall forecast correction has become an effective way and has received more and more attention [10-16]. For example, Shu Xingsheng et al. [10] used artificial neural network, extreme learning machine and support vector machine (SVM) models to conduct multi-model integrated forecast of rainfall in Huanren Reservoir basin in the next 1 to 3 days. The results showed that the effect of multi-model integration is better than that of a single model, and the SVM model has the most obvious improvement in rainfall forecast accuracy. Sun et al. [12] used the random forest (RF) model to correct the ERA5 precipitation data in 11 river basins in the upper reaches of the Tibetan Plateau from 1951 to 2020 and constructed a gridded daily precipitation dataset. The results showed that the corrected precipitation data are in good agreement with the observations. Nana et al. [14] applied machine learning algorithms to predict rainfall in different areas of Ghana. The results showed that RF and multilayer perceptron (MLP) performed well, while the *k*-nearest neighbor algorithm performed poorly.

This paper took the rainfall forecast correction of 12 stations in Chaobai River basin in the next 12 hours as the research object. Three kinds of correction models based on support vector machine, random forest and multilayer perceptron algorithms were constructed. The rainfall forecast data in different forecast periods were corrected and the correction results of each model were evaluated.

1 Model methods

1.1 Model principles

1.1.1 Random forest

Random forest (RF) regression is a bootstrap aggregation algorithm that uses decision trees as the base learner [17-18]. Based on bootstrap sampling of the dataset, and combined with the decision tree algorithm, different regression prediction models are created, and the final prediction results are obtained by comprehensively integrating different models. The RF algorithm includes the following main steps: the bootstrap technology is used to sample from the original dataset to generate multiple sub-datasets; for each sub-dataset, the random feature selection method is used to filter out feature variables for training the decision tree model; each decision tree is constructed using the CART algorithm, the best features is selected as split nodes, and the output value of the corresponding sample of the node on each leaf node is predicted; the prediction results of each decision tree are synthesized to obtain the final prediction result. The overall structure of the random forest model is shown in Figure 1.

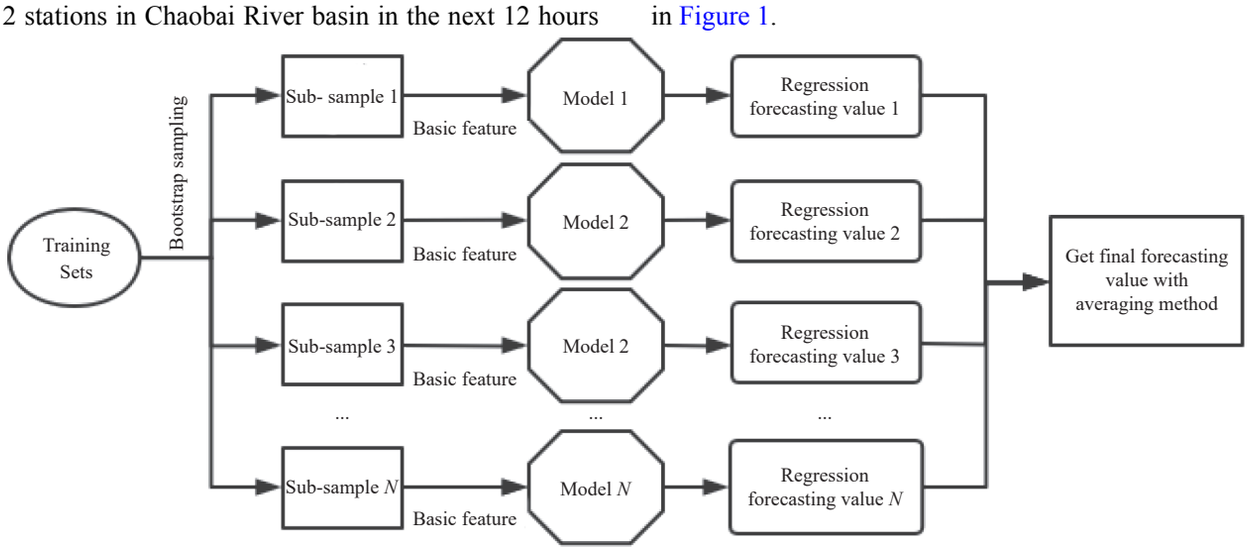


Fig. 1 Structure of random forest regression model

1.1.2 Multilayer perceptron

Multilayer perceptron (MLP) is a feedforward neural network, which consists of multiple neuron layers such as the input layer, the hidden layer and the output layer. The hidden layer can be multi-layered [19-20]. Each neuron is connected to all neurons in its previous layer and subsequent layer. Its structure is shown in Figure 2.

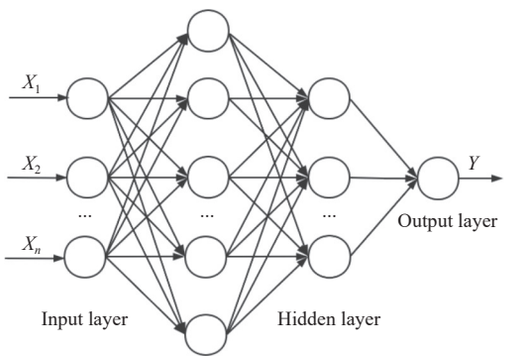


Fig. 2 Structure of multilayer perceptron model

When applying the MLP model to analyze regression problems, the output layer is usually a linear function that predicts the value of the target variable. Each neuron has a set of weights that map input values to the neuron's output. Each neuron also has a bias term, which is activated through activation functions such as Sigmod, Tanh, and ReLU. The model uses a loss function to measure the error between the predicted value and the actual value. Commonly used loss functions include mean square error and mean absolute error (MAE). Methods such as Adam optimizer and stochastic gradient descent algorithm are used to optimize model parameters to minimize the loss function, and the weight and bias of each neuron are updated through the backpropagation algorithm until a certain training accuracy is reached or the loss function is minimized. The general expression form of the MLP model is as follows

$$\boldsymbol{Y} = f(w_i \boldsymbol{X} + \boldsymbol{B}_i) \quad (1)$$

Where: $\boldsymbol{Y}$ is the output value; f is the activation function; w_i is the weight of the fully connected layer; $\boldsymbol{X}$ is the matrix of input layer variables; and $\boldsymbol{B}_i$ is the bias term.

1.1.3 Support vector machine

The basic idea of support vector machine (SVM) regression is to find a linear or nonlinear hyperplane such that the interval between this hyperplane and the data points is maximized, and to find a set of support vectors such that they are closest to the hyperplane and fall on the interval boundary [21-24]. In SVM, in order to minimize the error between the predicted value and the actual value and to have a low complexity of the model, a suitable kernel function is usually chosen to map the data from the input space to a high-dimensional feature space, and the optimal hyperplane is solved in the feature space. In linear regression models, SVM uses linear kernel functions, and in nonlinear regression, kernel functions such as radial basis functions are usually used. The general expression form of the SVM regression model is as follows

$$\min : \frac{1}{2}\|\omega\|^2 \quad (2)$$

$$\text{s.t.} \begin{cases} y_i - \omega x - b \varepsilon \\ \omega x + b - y_i \varepsilon \end{cases} \quad (3)$$

Where: ω is the weight vector; b is the bias term; y_i is the actual output value of the ith sample; ε is the interval size; and x is the set of optimal vectors to be solved. The objective is to minimize the squared norm of the weight vectors while ensuring that the prediction error of each sample does not exceed ε.

1.2 Model parameter estimation

Bayesian optimization technology is used to estimate the parameters of each of the above models. Compared with traditional hyperparameter optimization methods, Bayesian optimization technology takes into account the prior distribution of each hyperparameter, and obtains the best estimation of hyperparameters by continuously updating the prior distribution of each parameter [25-26]. Bayesian optimization technology samples the parameters in the region with higher likelihood and avoids traversing the entire hyperparameter space, thus having better estimation efficiency and accuracy. In addition, Bayesian optimization technology can conduct adaptive search in the hyperparameter space and find the global optimal solution more easily. On the basis of setting the value range of key hyperparameters in each model, efficient estimation of parameters can be carried out by Bayesian optimization technology. The setting of the value range

of key hyperparameters in the random forest model is shown in Table 1.

Tab. 1 Value range of some hyperparameters in the RF

n_estimators	max_ features	max_ depth	min_samples_ split	min_samples_ leaf
[10~100]	['sqrt', 'log2', auto']	[2~20]	[2~20]	[1~10]

In the table 1: n_estimators is the number of trees in the random forest; max_features is used to determine the range or specific value of the number of features used for splitting on each decision tree node; max_depth is the maximum depth of the decision tree; min_samples_split is the minimum number of samples required for splitting a decision tree node; and min_samples_leaf is the minimum number of samples required for leaf nodes. When the number of samples of a node is less than min_samples_leaf, the node will be pruned to become a leaf node. The setting of the value range of key hyperparameters in the multilayer perceptron model is shown in Table 2.

Tab. 2 Value range of some hyperparameters in the MLP

hidden_ layer_sizes	activation	alpha	learning_rate	solver	max_iter
[10~500]	['relu', 'tanh', 'logistic']	[0.000 001~ 100]	['constant', 'adaptive']	['adam']	[100~ 5 000]

Considering the small size of the dataset in this study, the MLP model with a single hidden layer is used to reduce the model complexity and risk of overfitting. Hidden_layer_sizes represents the number of neurons in the hidden layer; activation is the activation function; alpha is the L2 regularization coefficient; learning_rate is the learning rate; solver is the optimizer type; max_iter is the maximum number of iterations. The setting of the value range of key hyperparameters in the support vector machine model is shown in Table 3.

Tab. 3 Value range of some hyperparameters in the SVM

C	gamma	kernel	epsilon
Real(1e−6, 1e+4, 'log-uniform')	Real(1e−6, 1e+4, 'log-uniform')	['rbf']	Real(0, 1, 'uniform')

In the table 3: C is the regularization parameter, used to control the degree of punishment of the classifier on the misclassified samples; gamma is the kernel function parameter, used to control the distribution characteristics after sample points mapping to the high-dimensional space, Real(1e−6, 1e+4, 'log-uniform') defines a continuous real-type search space that ranges from 1e−6 to 1e+4, and the logarithm uniform distribution method is used to search in this range; kernel is the type of kernel function, and the radial basis function is chosen this time; epsilon is the floating-point numbers, that work on the loss function.

1.3 Model evaluation index

Root mean squared error (E_{RMS}) and coefficient of determination (R^2) are used to evaluate the accuracy of the original rainfall forecasts and the correction effect of each model on the forecasted rainfall in different forecast periods. The root mean squared error is calculated as follows

$$E_{\mathrm{RMS}} = \sqrt{\frac{\sum (y_{\mathrm{pred}} - y_{\mathrm{true}})^2}{N}} \tag{4}$$

Where: E_{RMS} is the root mean square error; N is the number of samples; y_{pred} is the predicted value of the model; y_{true} is the measured value. The smaller the value of E_{RMS} is, the closer the model's prediction results are to the actual measurement, and the higher the model accuracy is. The coefficient of determination is calculated as follows

$$R^2 = 1 - \frac{\sum (y_{\mathrm{pred}} - y_{\mathrm{true}})^2}{\sum (y_{\mathrm{true}} - \bar{y}_{\mathrm{true}})^2} \tag{5}$$

Where: R^2 is the coefficient of determination; y_{pred} is the predicted value of the model; y_{true} is the measured value; $\bar{y}_{\mathrm{true}}$ is the mean of the measured value. The closer R^2 is to 1, the higher the prediction accuracy of the model is.

2 Results analysis

2.1 Data information

This paper takes the forecast rainfall corrections of 12 stations in Chaobai River basin (Beijing city area) in the next 12 different forecast periods as the research object, and the data used is the hour-by-hour data from June to September of the 2021 and 2022 flood seasons, and the numerical forecast products will forecast the hour-by-hour rainfall of the next 12 hours in a rolling

manner on every hour, with a space resolution of 500 m×500 m.

2.2 Correcting effect analysis

The calibration models are constructed for each forecast period, that is, for the rainfall correction of each site in the 12 forecast periods, 12 different calibration models are required to construct. The model input is the forecast rainfall data of the corresponding grid of the station and its surrounding 8 grids (9 grids in total), that is, the forecast rainfall data from the 9 grids are used as the model input. The average value of the forecast rainfall from the 9 grids is used as the original forecast rainfall value corresponding to the station. 80% of the sample data are used for model training, and the remaining 20% of the sample data are used for model validation. Considering that the values of rainfall at different times vary greatly in magnitude, the rainfall data are normalized using the following equation.

$$Z = \frac{x - \mu}{\sigma} \tag{6}$$

Where: Z is the standardized variable; x is the input variable; μ is the mean of the input variable; σ is the standard deviation of the input variable.

The root mean square errors and coefficients of determination of the original forecast rainfall and the forecast rainfall corrected by each model in different forecast periods at each station in the rate period are shown in Figures 3 and 4, respectively. As can be seen from the figures, after correction by the machine learning model, the root mean square error values of the forecast rainfall data in different foresight periods at each station are reduced, while the values of coefficient of determination are improved, that is, each model has a better correction effect on rainfall in different forecast periods, which can improve the accuracy of the rainfall forecast, among which, the support vector machine has the best correction effect on data.

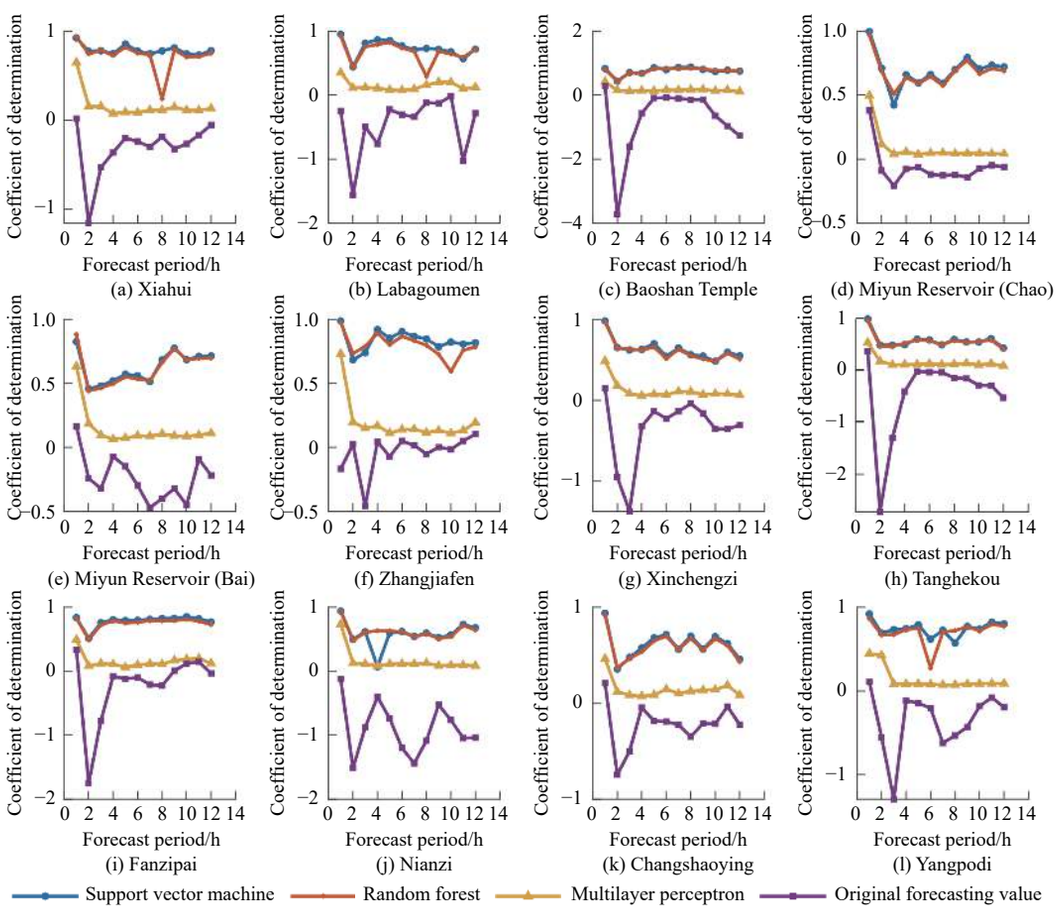


Fig. 3 R^2 at each station in train period

The mean values of root mean square error and coefficient of determination in different forecast periods at 12 stations in the rate period are shown in Table 4. As can be seen from the table, after correction with the 3 models, the coefficient of determination in each forecast period is significantly improved, and the

root mean square error is significantly reduced. Among the 3 models, support vector machine model is the best, followed by random forest model.

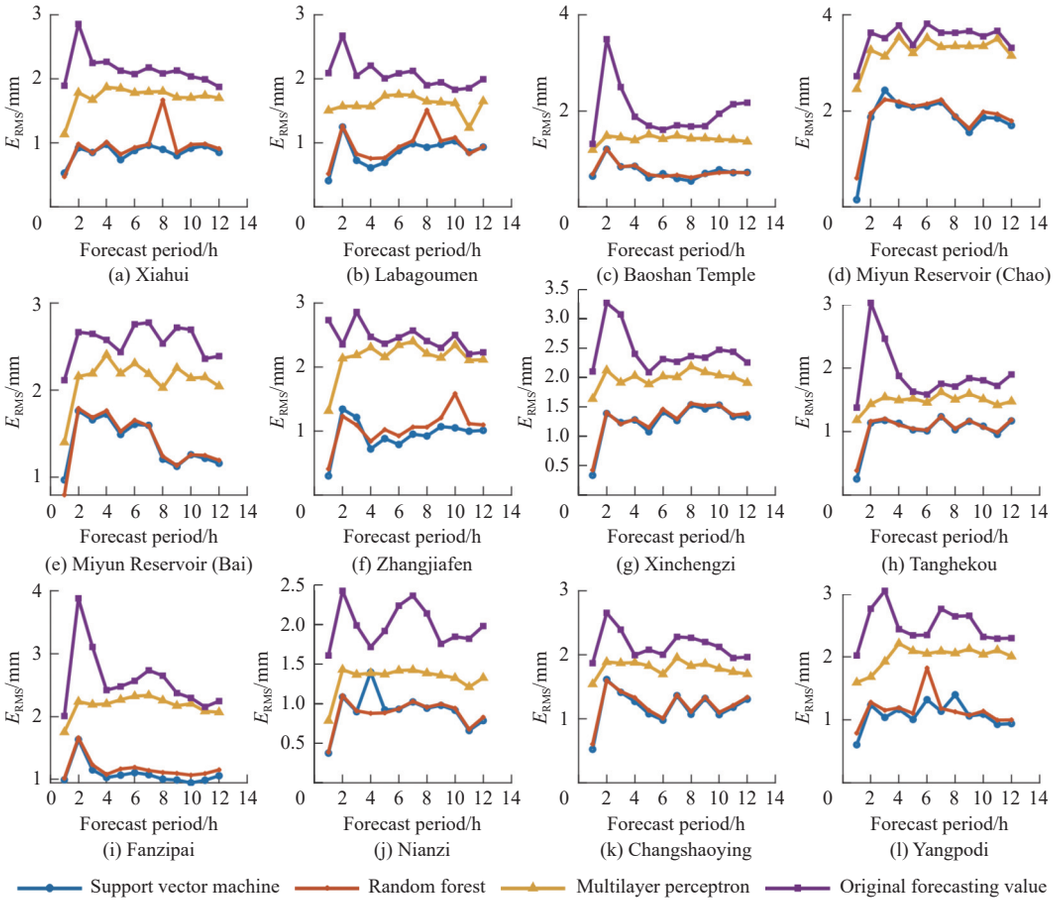


Fig. 4 E_{RMS} at each station in train period

Tab. 4 The mean value of E_{RMS} and R^2 of 12 stations in train period

Forecast period/h	R^2				E_{RMS}/mm			
	Original Values	Support vector machine	Random forest	Multilayer perceptron	Original Values	Support vector machine	Random forest	Multilayer perceptron
1	0.12	0.88	0.79	0.46	1.36	0.52	0.67	1.11
2	−1.29	0.56	0.51	0.16	2.32	1.08	1.18	1.60
3	−0.51	0.59	0.54	0.09	2.55	1.34	1.41	2.06
4	−0.14	0.63	0.58	0.07	1.82	0.99	1.10	1.65
5	−0.12	0.61	0.46	0.07	2.13	1.23	1.48	1.99
6	−0.37	0.56	0.40	0.06	2.09	1.21	1.43	1.77
7	−0.57	0.58	0.37	0.04	1.91	1.03	1.23	1.58
8	−0.34	0.48	0.28	0.05	2.17	1.33	1.61	1.87
9	−0.37	0.51	0.38	0.05	2.07	1.26	1.42	1.78
10	−0.36	0.54	0.40	0.06	2.14	1.25	1.42	1.85
11	−0.23	0.45	0.33	0.03	2.42	1.60	1.81	2.19
12	−0.13	0.44	0.36	0.11	2.40	1.71	1.83	2.16

The root mean square errors and coefficients of determination of the original forecast rainfall and the forecast rainfall corrected by each model in different forecast periods at each station in the validation period are shown in Figures 5 and 6, respectively. As can be seen from the figures, the conclusions reflected by each indicator are consistent with the rate period, that is, the accuracy of rainfall in different forecast periods is improved after correction with each model. Among the 3 models, support vector machine model is the best.

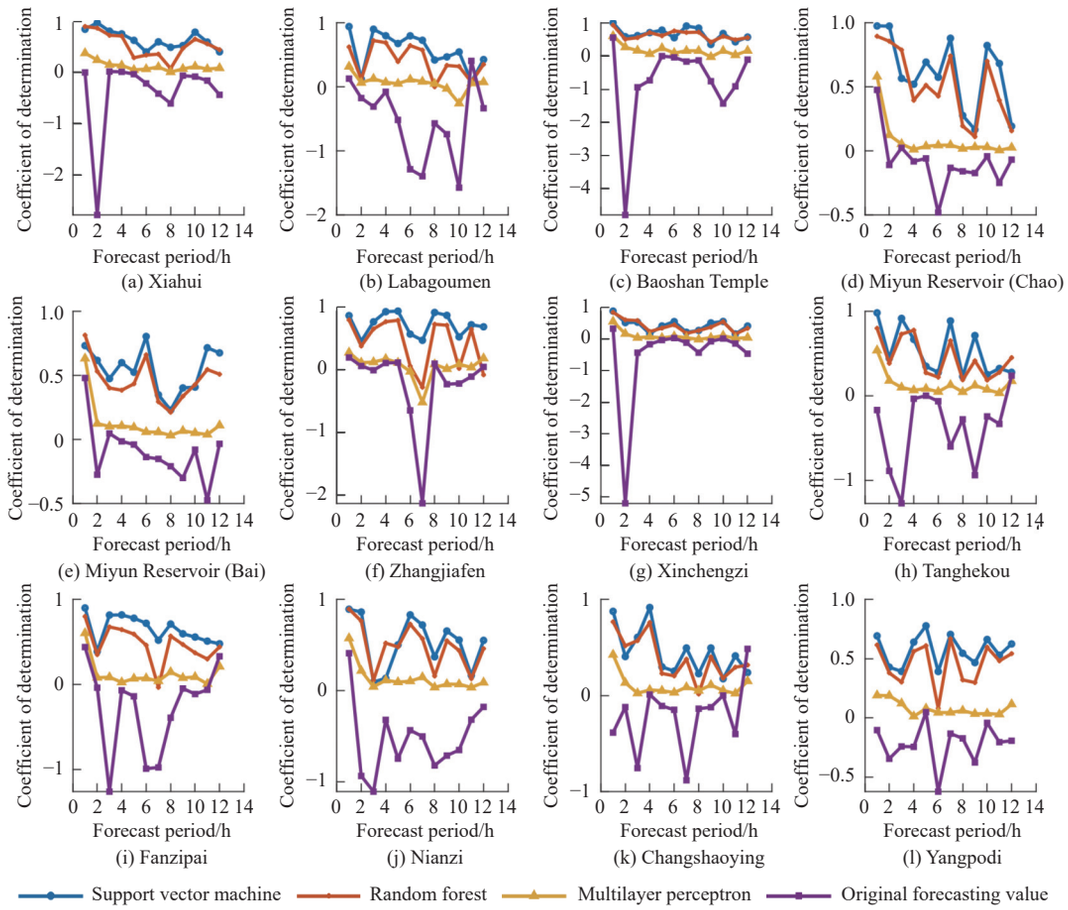


Fig. 5 R^2 at each station in test period

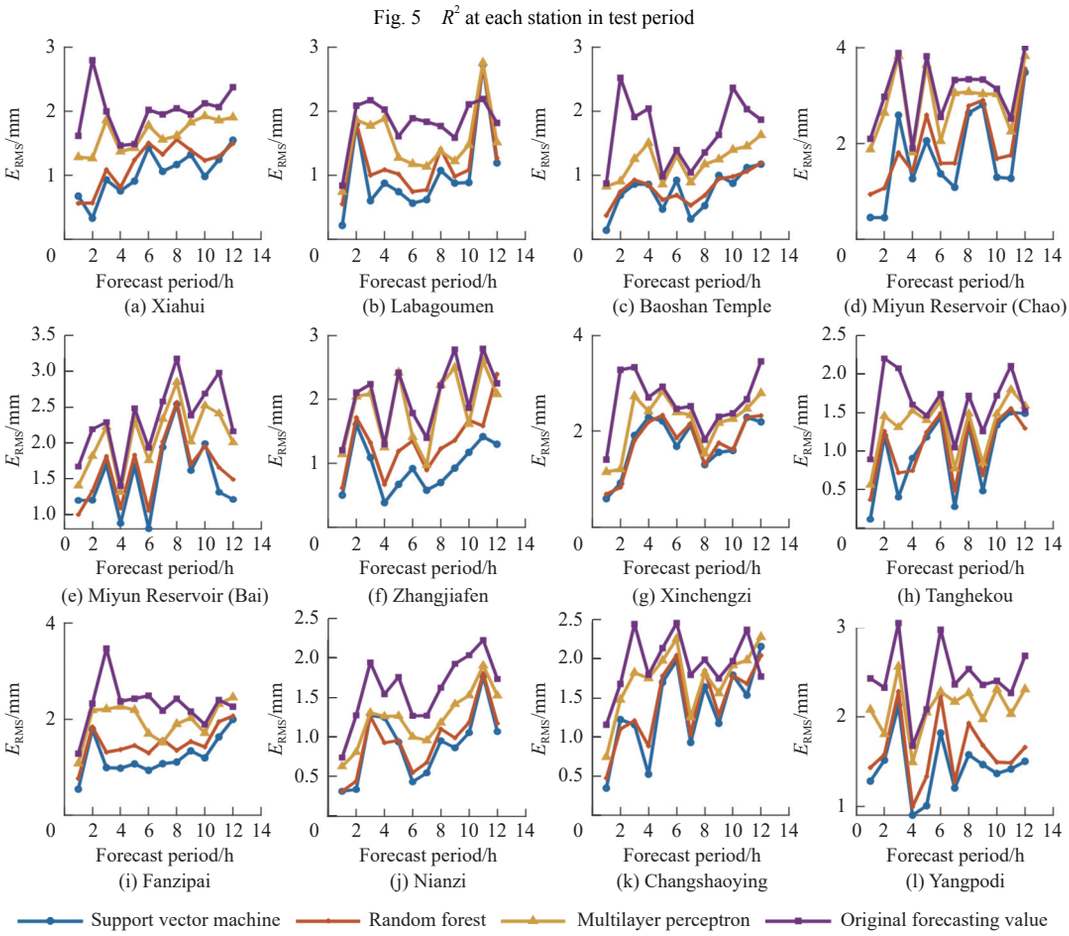


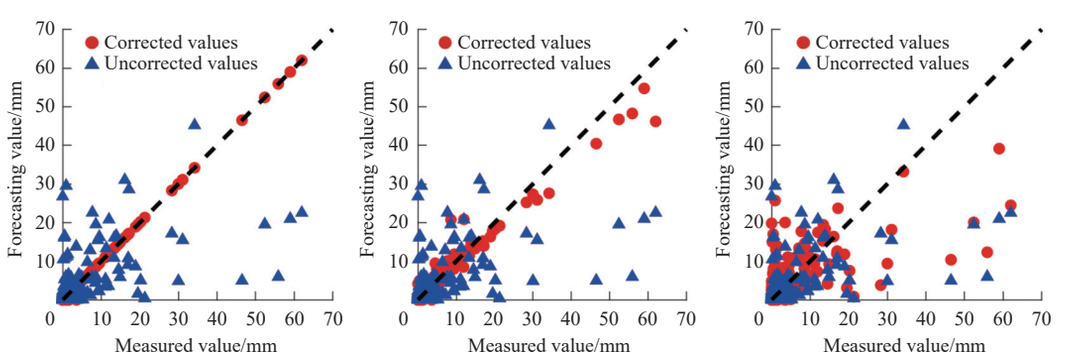
Fig. 6 E_{RMS} at each station in test period

The mean values of root mean square error and coefficient of determination in different forecast periods at 12 stations in the validation period are shown in Table 5. As can be seen from the table, the conclusion is consistent with the rate period. After correction with the 3 models, the coefficient of determination in each forecast period is significantly improved, and the root mean square error is significantly reduced. Among the 3 models, support vector machine model is the best, followed by random forest model.

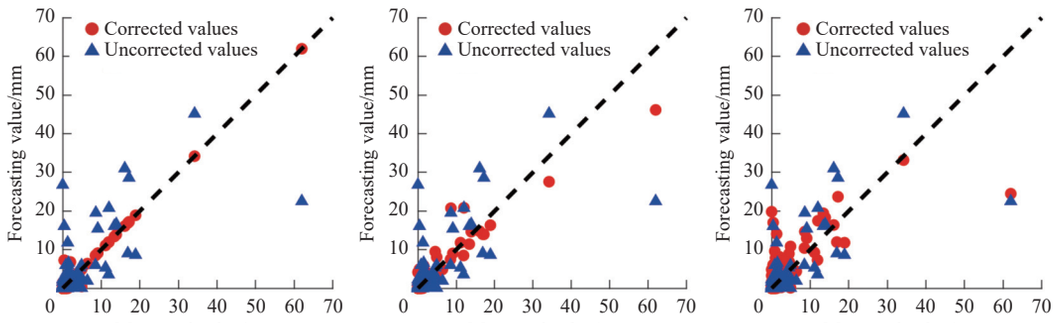
Tab. 5 The mean value of E_{RMS} and R^2 of 12 stations in test period

Forecast period/h	R^2				E_{RMS}/mm			
	Original Values	Support vector machine	Random forest	Multilayer perceptron	Original Values	Support vector machine	Random forest	Multilayer perceptron
1	0.12	0.88	0.79	0.46	1.36	0.52	0.67	1.11
2	−1.29	0.56	0.51	0.16	2.32	1.08	1.18	1.60
3	−0.51	0.59	0.54	0.09	2.55	1.34	1.41	2.06
4	−0.14	0.63	0.58	0.07	1.82	0.99	1.10	1.65
5	−0.12	0.61	0.46	0.07	2.13	1.23	1.48	1.99
6	−0.37	0.56	0.40	0.06	2.09	1.21	1.43	1.77
7	−0.57	0.58	0.37	0.04	1.91	1.03	1.23	1.58
8	−0.34	0.48	0.28	0.05	2.17	1.33	1.61	1.87
9	−0.37	0.51	0.38	0.05	2.07	1.26	1.42	1.78
10	−0.36	0.54	0.40	0.06	2.14	1.25	1.42	1.85
11	−0.23	0.45	0.33	0.03	2.42	1.60	1.81	2.19
12	−0.13	0.44	0.36	0.11	2.40	1.71	1.83	2.16

Taking the forecast rainfall of Miyun Reservoir (tide) in the first forecast period as an example, the scatter plots of the forecast rainfall and measured rainfall before and after model correction are drawn in Figure 7. As can be seen from Figure 7, the points between the uncorrected original forecast rainfall and the measured rainfall are more scattered, and after the model correction, the points of the forecast rainfall and the measured rainfall tend to be aggregated in the vicinity of the 1∶1 line, which indicates that the accuracy of the corrected forecast rainfall has been improved and is closer to that of the measured rainfall. It can also be seen from the figure that, the support vector machine model is the best both in the rate period and the validation period, followed by random forest model.



(a) Support vector machine in the rate period (b) Random forest in the rate period (c) Multilayer perceptron in the rate period



(d) Support vector machine in the validation period (e) Random forest in the validation period (f) Multilayer perceptron in the validation period

Fig. 7 Scattered plot of forecast rainfall and measured rainfall before and after model correction

The main reason why SVM model is better than the other two models is reflected in[27-29]: SVM is modeled with the principle of structural risk minimization, which minimizes the training error and reduces the generalization error at the same time, effectively avoids overfitting, and makes SVM have a strong generalization ability; SVM algorithm determines the decision boundary by maximizing the interval, which is robust to the outliers, and effectively avoids the influence of outliers on the results; the performance of SVM is affected by the number of training samples and sample distribution characteristics. In the case of small sample size, SVM can better deal with uneven distribution of data, and is more suitable for small sample data modeling.

3 Conclusion

Based on SVM, RF and MLP models, combined with Bayesian optimization technology, the correction models of forecast rainfall data in different forecast periods were constructed to correct and analyze the forecast rainfall data of 12 stations in the Chaobai River basin in 12 different forecast periods.

E_{RMS} and R^2 were used to evaluate the effect of the original forecast and the forecast corrected by the SVM, RF and MLP models. In terms of the average values of R^2 corresponding to the 12 forecast periods, they are −0.37, 0.69, 0.67 and 0.15 respectively in the rate periods, and they are −0.36, 0.57, 0.45 and 0.11 respectively in the validation period. Each correction model has a good correction effect on the forecast rainfall in different forecast periods at each station; in terms of the E_{RMS} index, after correction by SVM, RF and MLP models, the average value of E_{RMS} decreased by 54.2%, 50.0% and 20.8% respectively in the rate period; the reduction was 42.9%, 33.3% and 14.3% respectively in the verification period. Among the 3 correction models, SVM model is the best, followed by RF model.

Compared with parameter optimization methods commonly used in machine learning such as grid search and random search, the Bayesian optimization method adopted this time can obtain the optimal solution of parameters under a relatively small operating load, and can obtain the probability distribution estimation of parameters to analyze the uncertainty of parameter estimation. In addition, the probability distribution estimation based on parameters, combined with Markov chain Monte Carlo sampling technology, can obtain multiple sets of model parameters, which can then be used to analyze the uncertainty of rainfall correction. In the follow-up work, this topic will be further studied.

References:

[1] HUANG Y X, WANG Q Z, LIANG Z M, et al. Research advances on real-time correction methods for flood forecasting[J]. South-to-North Water Transfers and Water Science & Technology (Chinese and English), 2021, 19(1): 12-35. DOI: 10.13476/j.cnki.nsbdqk.2021.0002.

[2] WEI J H, HUANG Y C, YAO C. Adaptability assessment of precipitation forecast products in small and medium-sized basins under different hydrometeorological divisions[J]. South-to-North Water Transfers and Water Science & Technology (Chinese and En-

glish), 2022, 20(6): 1208-1219. DOI: 10.13476/j.cnki.nsbdqk.2022.0119.

[3] DI S C, LI Z M, LIU Y, et al. Research of approaching rainfall forecast method based on weather radar inversion and cloud image extrapolation[J]. Water Resources and Hydropower Engineering (Chinese and English), 2022, 53(5): 13-21. DOI: 10.13928/j.cnki.wrahe.2022.05.002.

[4] LIU Z Y, LIU Y H, KONG X Y. Problems, strategies and key technology research of flood forecasting and early warning for small and medium-sized rivers[J]. Journal of Hohai University (Natural Sciences), 2021, 49(1): 1-6. DOI: 10.3876/j.issn.1000-1980.2021.01.001.

[5] ZHANG Y L, ZHANG W G, JIA B Y, et al. Rainfall forecast accuracy evaluation method for flood control demand[J]. South-to-North Water Transfers and Water Science & Technology (Chinese and English), 2021, 19(2): 293-300. DOI: 10.13476/j.cnki.nsbdqk.2021.0031.

[6] HU Y M, LIANG Z M, JIANG X L, et al. Study on statistical post-processing of GFS ensemble precipitation forecasts[J]. South-to-North Water Transfers and Water Science & Technology (Chinese and English), 2019, 17(1): 15-19. DOI: 10.13476/j.cnki.nsbdqk.2019.0003.

[7] TANG R, WANG Y T, LI M, et al. Accuracy evaluation of ECMWF precipitation forecast in different utilization forms[J]. China Rural Water and Hydropower, 2020, 453(7): 1-5. DOI: 10.3969/j.issn.1007-2284.2020.07.001.

[8] WEN L C, LI Z J. Application of the corrected numerical rainfall in flood forecasting[J]. Water Resources and Power, 2010, 28(4): 1-4. DOI: 10.3969/j.issn.1000-7709.2010.04.001.

[9] WU X S, WANG Z L, CHEN K B, et al. A precipitation combined forecasting model based on atmospheric circulation and sea surface temperature[J]. Water Resources Protection, 2022, 38(6): 81-87. DOI: 10.3880/j.issn.1004-6933.2022.06.011.

[10] SHU X S, WANG Z R, LI F W, et al. Short-term rainfall multi-mode integrated forecasting based on machine learning models[J]. South-to-North Water Transfers and Water Science & Technology (Chinese and English), 2020, 18(1): 42-50. DOI: 10.13476/j.cnki.nsbdqk.2020.0006.

[11] ORTIZ-GARCIA E G, SALCEDO-SANZ S, CASANOVA-MATEO C. Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data[J]. Atmospheric Research, 2014, 139: 128-136. DOI: 10.1016/j.atmosres.2014.01.012.

[12] SUN H, YAO T D, SU F G, et al. Corrected ERA5 precipitation by machine learning significantly improved flow simulations for the Third Pole basins[J]. Journal of Hydrometeorology, 2022, 23(10): 1663-1679. DOI: 10.1175/JHM-D-22-0015.1.

[13] ADARYANI F R, MOUSAVI S J, JAFARI F. Short-term rainfall forecasting using machine learning-based approaches of PSO-SVR, LSTM and CNN[J]. Journal of Hydrology, 2022, 614: 128463. DOI: 10.1016/j.jhydrol.2022.128463.

[14] APPIAH-BADU N K A, MISSAH Y M, AMEKUDZI L K, et al. Rainfall prediction using machine learning algorithms for the various ecological zones of Ghana[J]. IEEE Access, 2021, 10: 5069-5082. DOI: 10.1109/ACCESS.2021.3139312.

[15] KO C M, JEONG Y Y, LEE Y M, et al. The development of a quantitative precipitation forecast correction technique based on machine learning for hydrological applications[J]. Atmosphere, 2020, 11(1): 111. DOI: 10.3390/atmos11010111.

[16] LI H Y, ZHANG Y, LEI H J, et al. Machine learning-based bias correction of precipitation measurements at high altitude[J]. Remote Sensing, 2023, 15(8): 2180. DOI: 10.3390/rs15082180.

[17] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45: 5-32. DOI: 10.1023/A:1010933404324.

[18] ZOUNEMAT-KERMANI M, BATELAAN O, FADAEE M, et al. Ensemble machine learning paradigms in hydrology: A review[J]. Journal of Hydrology, 2021, 598: 126266. DOI: 10.1016/j.jhydrol.2021.126266.

[19] CHOUBIN B, KHALIGHI-SIGAROODI S, MALEKIAN A, et al. Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals[J]. Hydrological Sciences Journal, 2016, 61(6): 1001-1009. DOI: 10.1080/02626667.2014.966721.

(下转第 950 页)