DOI:10.13870/j.cnki.stbcxb.2025.06.018 CSTR:32310.14.stbcxb.2025.06.018
王晓燕,刘刚,刘依嘉,等.基于机器学习模型的塔里木河流域卫星降水融合校正[J].水土保持学报,2025,39(6):419-430.
WANG Xiaoyan, LIU Gang, LIU Yijia, et al. Correction of satellite precipitation product based on machine learning models in Tarim River basin [J]. Journal of Soil and Water Conservation, 2025, 39(6):419-430.

基于机器学习模型的塔里木河流域卫星降水融合校正

王晓燕[1,2], 刘 刚[1,2], 刘依嘉[1,2], 王杰斌[1,2], 吴仁达[3], 谷黄河[1,2]
(1.河海大学水灾害防御全国重点实验室,南京 210098; 2.河海大学水文水资源学院,南京 210098;
3.中国电建集团成都勘测设计研究院有限公司,成都 610072)

摘 要:[**目的**]为识别多源数据融合方法在区域高精度降水数据构建中的适用性。[**方法**]以塔里木河流域为研究区,对比了4种机器学习方法(随机森林、支持向量机、XGBoost及回归树)对区域多源融合降水模拟的效果差异,探讨不同卫星降水融合源(GPM IMERG-v06(简称GPMv06)和GPM IMERG-v07(简称GPMv07))及考虑NDVI对降水响应的滞后性对融合模型精度的影响。[**结果**]GPMv06和GPMv07均呈高估低海拔降水、低估高海拔降水的特征。与GPMv06相比,GPMv07夏冬季节的降水精度均有提升,尤其冬季降水的纳什效率系数提高0.58。4种融合模型中XGBoost方法的月降水精度最高。与GPMv07相比,XGBoost方法得到的融合月降水均方根误差减小2.01 mm,纳什系数不低于0.6的站点占比及纳什系数平均值分别提升33%和0.23。卫星降水输入误差对降水融合模型XGBoost的精度影响较小。在春夏季节的部分月份考虑NDVI对降水的滞后性有利于提升XGBoost融合模型的精度,而考虑NDVI对降水的滞后性对秋冬季节多数月份融合降水精度的影响较小。[**结论**]XGBoost方法在塔里木河流域GPMv07卫星降水校正中有较大优势,校正后的卫星降水精度有较大提升。研究结果可为区域水资源管理及水土流失预防研究提供数据参考。

关键词:卫星降水; 融合校正; 机器学习; 塔里木河流域; GPM IMERG
中图分类号:TP79;P333 **文献标识码:**A **文章编号:**1009-2242(2025)06-0419-12

Correction of Satellite Precipitation Product Based on Machine Learning Models in Tarim River Basin

WANG Xiaoyan[1,2], LIU Gang[1,2], LIU Yijia[1,2], WANG Jiebin[1,2], WU Renda[3], GU Huanghe[1,2]
(1.*The National Key Laboratory of Water Disaster Prevention*, *Hohai University*, *Nanjing* 210098, *China*;
2.*College of Hydrology and Water Resources*, *Hohai University*, *Nanjing* 210098, *China*; 3.*Power China Chengdu Engineering Corporation Limited Chengdu* 610072, *China*)

Abstract: [**Objective**] To assess the applicability of a multi-source data fusion approach for constructing high-precision regional precipitation datasets. [**Methods**] Taking the Tarim River basin as the study area, the study compared the performance of four machine learning methods (random forest, support vector machine, XGBoost, and regression tree) in simulating regional precipitation using multi-source data fusion. It further discussed the effects of different satellite precipitation products (GPM IMERG-v06 (GPMv06) and GPM IMERG-v07 (GPMv07)) and incorporating the lagged response of the normalized difference vegetation index (NDVI) to precipitation on fusion model accuracy. [**Results**] Both GPMv06 and GPMv07 overestimated the precipitation in low-elevation zones and underestimated it in high-elevation zones. Compared with GPMv06, GPMv07 demonstrated improved precipitation prediction accuracy in both summer and winter, with the Nash-Sutcliffe efficiency (NSE) coefficient for winter precipitation increasing by 0.58 in particular. Among the four fusion models, XGBoost model achieved the highest accuracy in monthly precipitation simulation. Compared to

收稿日期:2025-04-14 **修回日期:**2025-06-04 **录用日期:**2025-06-24 **网络首发日期**(www.cnki.net):2025-08-29
资助项目:国家自然科学基金项目(42277074);中央高校基本科研业务费项目(B240201075)
第一作者:王晓燕(1986—),女,博士,副教授,主要从事水文及水资源研究。E-mail:xywang@hhu.edu.cn
通信作者:谷黄河(1986—),男,博士,副教授,主要从事水文及水资源研究。E-mail:ghh0001@hhu.edu.cn

GPMv07, the XGBoost model reduced the root-mean-square error (RMSE) by 2.01 mm, increased the percentage of sites with NSE coefficient no less than 0.6 by 33%, and raised the mean NSE coefficient by 0.23. The input error of satellite precipitation had a relatively minor influence on the accuracy of XGBoost model. Incorporating the lagged response of NDVI to precipitation improved the model's accuracy in some spring and summer months but had limited effect for most autumn and winter months. [**Conclusion**] The XGBoost model demonstrates significant advantages in correcting the GPMv07 satellite precipitation data in the Tarim River basin, achieving substantial improvements in satellite precipitation prediction accuracy. This research provides data references for studies in regional water resource management and soil erosion prevention.

Keywords: satellite precipitation; fusion correction; machine learning; Tarim River basin; GPM IMERGE

Received: 2025-04-14 **Revised**: 2025-06-04 **Accepted**: 2025-06-24 **Online**(www.cnki.net): 2025-08-29

降水在全球水循环、气候系统和生态系统中扮演着重要的角色[1]。由于降水具有较高的空间异质性,获取高质量的降水数据极具挑战性[2]。目前常用的降水估算方法包括站点观测、雷达观测及遥感观测[3]。站点实测数据被认为是降水真值,但当站点稀少或空间分布不均时,观测降水难以描绘流域降水的空间分布,卫星遥感数据由于其可以大范围地监测,已成为研究降水空间分布特征的重要手段[4-5]。但受反演算法、气象环境等的影响,卫星遥感产品存在随机误差和系统误差[6-7]。因此,如何获取高质量的降水产品已成为当前研究的热点和难点。

降水数据融合是用于整合不同类型降水数据优势的常用方法[8],可对区域降水特征形成全面的刻画。近年来,国内外学者已进行了大量星地降水数据融合的研究,卢新玉等[9]利用概率密度匹配与最优插值2步融合校正方法成功提高新疆地区GPM IMERG卫星降水数据的精度;潘旸等[10]基于贝叶斯融合方法对高分辨率的雷达估测降水、卫星反演降水与地面站点观测降水3种降水数据进行融合,得到的多源融合产品精度优于任意单一来源的降水产品。以上研究仅聚焦于降水数据本身,未考虑地形、空间位置等信息对降水融合精度的提升。基于此,石羽佳等[11]基于站点观测数据、卫星降水数据(GPM)及地理数据,结合人工神经网络方法进行多源数据融合校正,生成能更好捕捉降水空间分布的海河流域融合降水数据集;阮惠华等[12]提出了一种基于XGBoost算法和地统计学理论的地面观测-雷达-卫星遥感多源降水融合方法,通过考虑降水的时间相关性,显著提高广东省北部山区逐时降水数据的精度;在天山地区,LU等[13]采用逐步回归、地理加权回归及随机森林3种方法对GPM IMERG降水产品进行校正发现,3种方法均可提高产品精度,但在低海拔区域随机森林法表现最佳。相较于传统的回归方法,机器学习方法因具有处理非线性问题的能力、融合精度较高等优势,已被广大学者[11-13]关注和使用。然而,目前的星地降水数据融合研究仍存在一定的不足,从数据方面,针对卫星数据GPM IMERG(v06)融合校正方面的研究成果较多,而最新一代GPM IMERG(v07)卫星降水于2023—2024年陆续发布,目前基于GPM IMERG v07的降水融合研究较少,基于GPM IMERG v06和v07版本的降水融合模型的精度差异尚未揭示。从方法方面,已有研究[14]指出,NDVI是降水的重要影响因素,且NDVI可能对降水的响应存在滞后效益,但这种滞后性是否对降水融合模型精度产生影响的研究较缺乏。

西北地区是中国干旱、半干旱范围最大的地区,也是"一带一路"倡议的关键节点。降水在西北地区的生态环境和社会经济发展中起着重要作用。但西北地区地形地貌复杂多样,降水量空间差异大、局地性强,站点稀疏,难以获取准确的降水空间数据[15],阻碍区域水资源管理和生态环境保护等领域的精细化研究。鉴于此,本文选择西北地区典型流域塔里木河流域为研究区,基于GPM IMERG v06和v07及站点观测降水、地理时空数据等,采用随机森林、支持向量机、回归树及XGBoost 4种机器学习模型进行多源数据融合,对比分析不同融合方法的效果差异,探讨不同版本卫星降水数据及考虑NDVI对降水响应的滞后性对融合方法精度的影响,研究成果可为塔里木河流域水资源管理和水土保持预防研究提供可靠的降水数据基础。

1 研究区域与数据来源

1.1 研究区概况

塔里木河流域(图1)是中国面积最大的内陆河流域,位于新疆维吾尔自治区南部。整个流域由塔里木盆地的九大水系、塔里木河干流、塔克拉玛干沙漠和库姆塔格沙漠组成。塔里木河源自喀喇昆仑山,沿塔克拉玛干沙漠的北缘流向罗布泊,总长度达到2 327 km。流域内的水资源主要依赖于四周山区的冰

雪融水和降水,每年的径流量约为398.3亿m³[16]。塔里木河流域属于典型的干旱大陆性气候,具有降水稀少、蒸发强烈、风沙浮尘多和日照时间长等特征。多年平均气温为12.3 ℃,年平均气温最高为12.9 ℃,最低为11.5 ℃。多年平均降水量为68 mm[17]。

1.2　数据来源

1.2.1　*地面观测数据*　本研究所使用的观测降水数据包括塔里木河流域2001—2017年25个气象站和2001—2010年5个水文站的逐月降水数据,前者来源于中国气象局国家气候中心气象数据网(https://data.cma.cn/),后者来源于新疆塔里木河流域管理局。站点空间分布见图1。

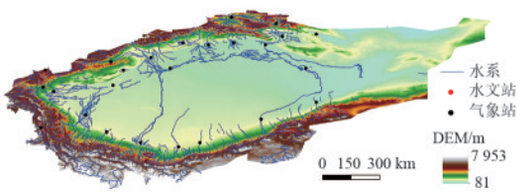


图1　塔里木河流域概况

Fig. 1　Overview of Tarim River Basin

1.2.2　*卫星降水产品*　全球降水测量(global precipitation measurement,简称GPM)是由美国国家航空航天局(NASA)和日本宇宙航空研究开发机构(JAXA)联合发起的一项国际卫星任务,旨在提供高精度、高分辨率的全球降水数据[18]。本研究使用2种GPM降水产品,分别为Integrated Multi-satellitE Retrievals for Global Precipitation Measurement Final Run v06(GPM IMERG-v06)和Integrated Multi-satellitE Retrievals for Global Precipitation Measurement Final Run v07(GPM IMERG-v07),时间为2001—2017年,时间分辨率为月,空间分辨率为0.1°×0.1°,数据来源于美国国家航空航天局官方网站(https://pmm.nasa.gov/)。下文中GPM IMERG-v06和GPM IMERG-v07分别简称为GPMv06和GPMv07。

1.2.3　*其他数据*　由于海拔、坡度、NDVI等都与降水有着密切的联系[19],本文使用的其他数据包括DEM数据和NDVI数据,DEM数据(90 m)来源于地理空间数据云(https://www.gscloud.cn),用于制作流域的坡度、坡向数据。NDVI月数据来源于国家青藏高原科学数据中心(https://data.tpdc.ac.cn/home),空间分辨率为250 m,时间为2001—2017年。这些数据均作为降水融合模型的输入数据。

2　研究方法

2.1　机器学习方法

1)支持向量回归法

支持向量回归(support vector regression,简称SVR):是一种用于回归分析的机器学习模型,是支持向量机(SVM)的扩展。SVR的目标是找到一个回归函数$f(x)$,使大多数数据点落在该函数的不敏感带内,同时尽量保持模型的复杂度(使权重向量的范数最小)。

$$f(x)=\langle w,x\rangle+b \tag{1}$$

式中:w为权重向量;b为偏置项;x为自变量。

SVR能够处理非线性回归问题,在小样本数据集中表现良好[20]。然而,对于大规模数据集,SVR的训练时间较长。

2)回归树法

回归树(regression tree,简称RT):是一种用于回归分析的决策树模型。它通过将数据集划分成多个区域,并在每个区域内拟合一个简单的模型(通常是常数值),来进行预测。回归树的目标是通过递归分割数据空间,找到最优的分割点,以最小化每个区域内的预测误差。该模型能够处理非线性关系和高维数据,易于理解和解释,结构直观,且对缺失值和噪声数据具有鲁棒性[21]。

3)随机森林法

随机森林法(random forest,简称RF):是一种集成学习方法,通过构建多个决策树并结合其预测结果来提高模型的准确性和稳定性,适用于分类和回归任务[22]。随机森林法通过引入随机性生成多样化的树,从而减少过拟合并提高泛化能力。在随机森林中,使用引导法(Bootstrap)从原始数据集中抽取多个子集,每个子集都用来训练一棵决策树。为了增加树的多样性,在每棵树的节点分裂过程中,只考虑部分特征。这种方法被称为特征随机性。具体地,对于每一棵决策树,执行的步骤为:①从原始数据集中有放回地随机抽取n个样本,构建训练集;②在每个节点分裂时,随机选择m个特征($m\leqslant d$,其中d是总特征数);③选择这m个特征中能最大化分裂效果的特征进行节点分裂。

回归任务中,随机森林法的预测值是所有树预测值的平均值:

$$\hat{y}=\frac{1}{T}\sum_{t=1}^{T}f_t(x) \tag{2}$$

式中:T为决策树的数量;$f_t(x)$为第t棵树的预测值。

随机森林法的优点包括能够处理高维数据和大量特征,对缺失值和噪声数据具有鲁棒性,通过集成多个树减少单一决策树的过拟合问题,且具有较高的准确性和稳定性。

4)XGBoost法

XGBoost(extreme gradient boosting,简称XGB):是一种基于梯度提升(gradient boosting)的集成学习方法,专为提高模型性能和效率而设计[23]。它通过

构建多个弱学习器(通常是决策树),逐步优化模型的预测误差,实现高精度的预测。XGBoost在许多机器学习竞赛中表现优异,广泛应用于分类和回归任务。在XGBoost中,目标是最小化损失函数:

$$L(\theta)=\sum_{i=1}^{n} l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \quad (3)$$

式中:l为损失函数;Ω为正则化项,用于控制模型复杂度;f_k为第k个弱学习器;$\hat{y}_i$为模型的预测值。

XGBoost在构建决策树时,引入了列抽样(column subsampling)和行抽样(row subsampling)的方法,进一步增加模型的随机性和稳定性。该方法提高计算效率并减少过拟合。其优点包括:高效的计算性能,支持并行计算,通过正则化项控制模型复杂度,减少过拟合,支持处理缺失值和噪声数据,提供特征重要性分析,便于解释模型[24]。

2.2 融合模型的构建及评价

本研究结合不同机器学习方法,考虑气象要素、地形要素、植被因子等,构建降水融合模型。研究采用点对网格匹配的方法,融合模型以所有站点降水数据为目标值,输入变量包括降水站点相应位置栅格的卫星降水量、NDVI值、高程、坡度、坡向、经度和纬度,利用机器学习方法进行关系挖掘。考虑到不同输入数据空间分辨率的差异,文中统一将各输入数据的空间分辨率转化为0.1°×0.1°。值得注意的是,由于水文站点的观测降水数据时段为2001—2010年,则站点相应位置栅格的卫星降水量、NDVI值的时段也仅取2001—2010年。而气象站观测降水及站点相应网格的卫星降水量、NDVI值的时段均取2001—2017年。模型最优参数采用粒子群优化算法[25]进行确定。为定量评估构建的降水融合模型的性能,使用空间5折交叉验证法用于参数验证,最终以站点降水观测为基准进行站点降水融合模型的评估。评估标准选择相对误差(BIAS)、均方根误差(RMSE)、相关系数(CC)及纳什效率系数(NSE)。

3 结果与分析

3.1 GPM降水精度的评估

GPMv06和GPMv07在不同季节降水精度评估指标见表1,可见不同版本的GPM均可以再现多数季节降水的平均值和变化趋势,冬季外的其他季节降水相关系数均达到0.65以上,相对偏差的绝对值低于0.15。但冬季降水与观测降水的相关系数低于0.4,相对偏差为−0.43～−0.37。GPMv07夏季和冬季降水的精度高于GPMv06,均方根误差均分别减小4.42、2.08 mm,纳什效率系数分别增大0.12和0.58;二者在春秋季节的均方根误差和纳什效率系数相近。总之,不同版本的GPM对冬季降水的平均值和变化趋势的再现能力低于其他季节。与GPMv06相比,GPMv07夏冬季节的降水精度均有提升,尤其冬季的提升幅度更显著,纳什效率系数提高0.58,但仍存在冬季降水量显著低估的不足。

表1 GPM卫星降水数据逐季节精度

Table 1 Seasonal accuracy of GPM satellite precipitation data

季节	CC		BIAS		RMSE/mm		NSE	
	GPMv06	GPMv07	GPMv06	GPMv07	GPMv06	GPMv07	GPMv06	GPMv07
春季	0.70	0.70	0.04	0.08	17.06	17.03	0.47	0.47
夏季	0.67	0.75	0.12	0.13	40.52	36.10	0.41	0.53
秋季	0.75	0.76	0.10	−0.09	15.20	15.07	0.56	0.57
冬季	0.31	0.37	−0.37	−0.43	10.56	8.48	−0.63	−0.05

图2为GPMv06和GPMv07在不同月份降水精度评估指标,不同版本GPM在5—9月的月降水精度高于其他月份,相关系数达到0.65以上,相对误差的绝对值在0.20以内,纳什效率系数高于0.40。与GPMv06相比,GPMv07月降水的精度在多数月份均有提升,相关系数和纳什系数提升最高的月份均为1月,分别提升0.10和0.48;均方根误差减幅最大的为8月。但在10—12月,GPMv07月降水精度略低于GPMv06。可见,不同版本的GPM在多雨期(5—9月)的降水精度高于其他月份,与GPMv06相比,GPMv07显著提升多数冬季月份的降水精度,尤其对于1月降水精度的提升幅度最高。

为进一步评估GPM降水在不同海拔区的精度差异,以2 000 m作为阈值,将站点分为高海拔和低海拔站点2类。图3为GPM在不同海拔区月降水的性能评估。总体来看,不同版本GPM均呈现高估低海拔降水、低估高海拔降水的特征,尤其显著高估低海拔区降水量,平均相对误差为0.29～0.34。结合各统计指标可知,GPM低海拔月降水的精度高于高海拔区,呈现较高的相关系数(>0.8)和较小的均方根误差(7.60～8.60 mm)。例外的是,GPMv06在低海拔的纳什效率系数平均值低于高海拔站点,可能与GPMv06在低海拔处有更大的相对偏差(>0.20)有关。与GPMv06相比,GPMv07在不同海拔区的精度均有提升,如均方根误差在高低海拔区分别减小0.40、1.00 mm,纳什效率系数在高海拔区提升约0.1。

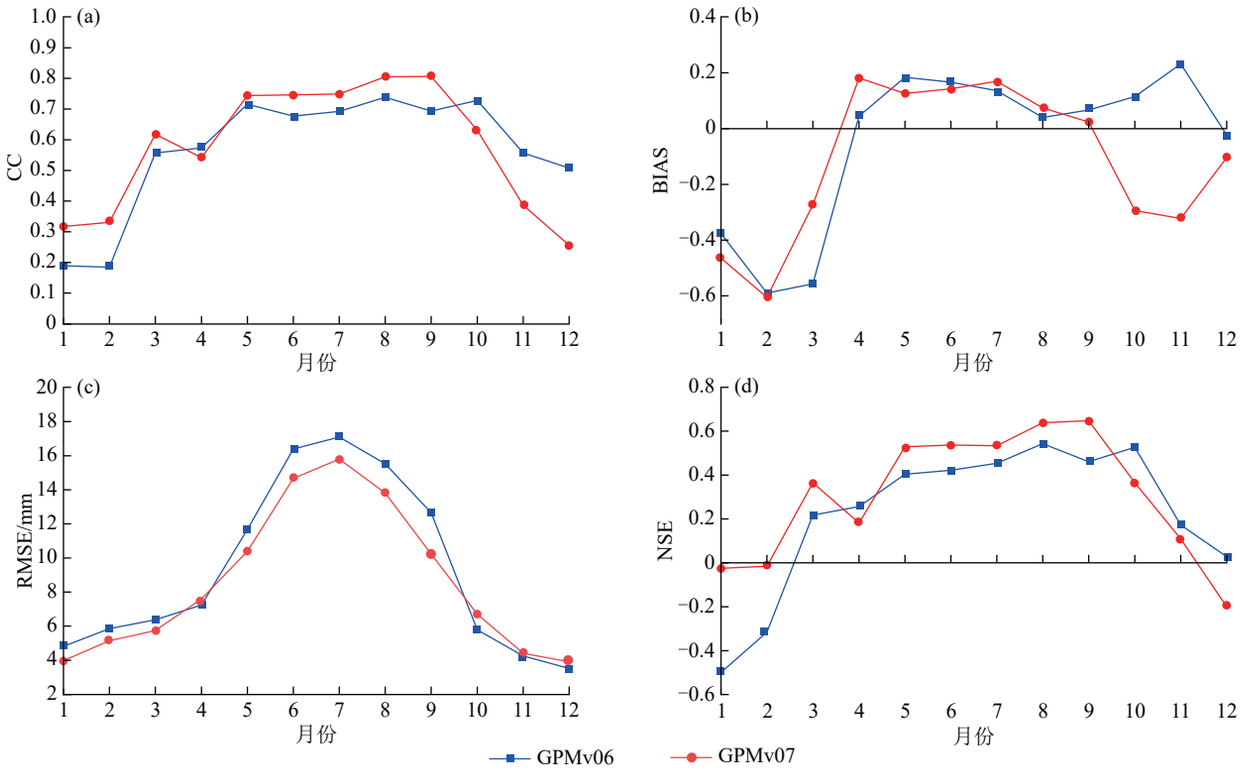
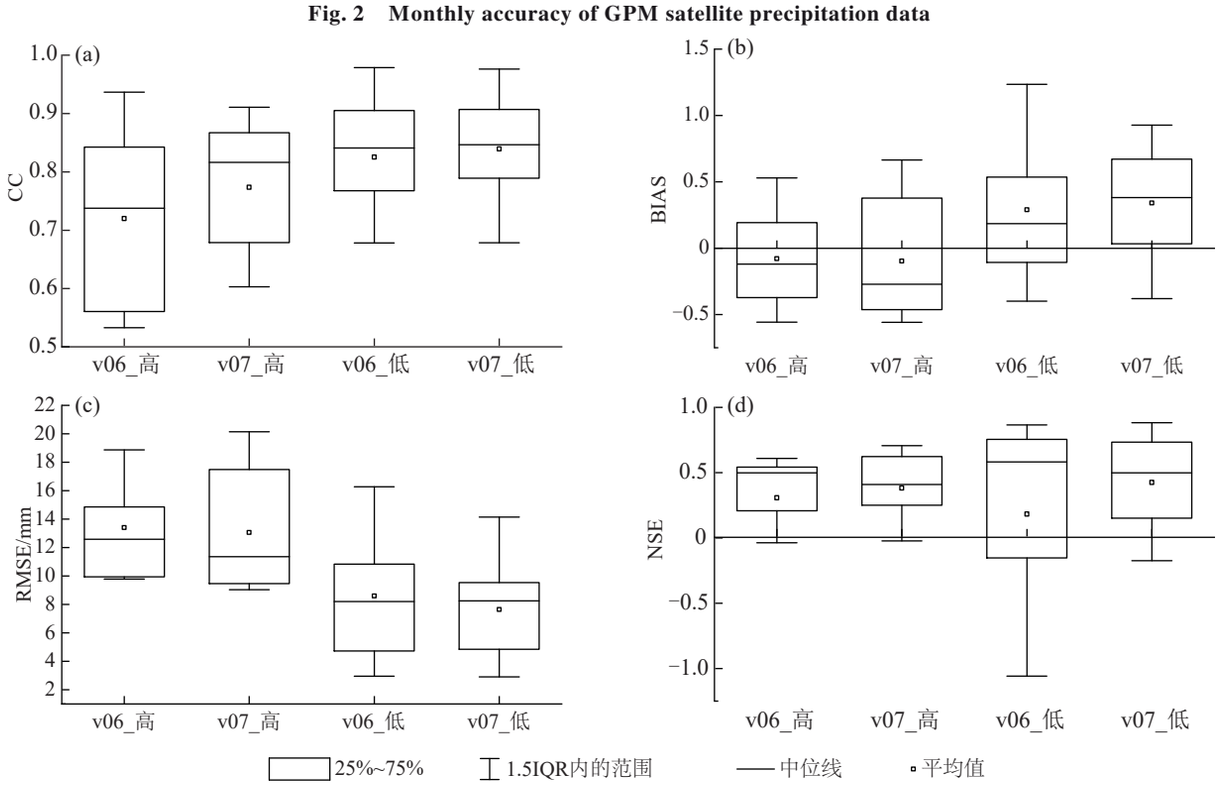


图 2　GPM 卫星降水数据逐月精度

Fig. 2　Monthly accuracy of GPM satellite precipitation data



注:图例中的 IQR 为 75% 分位数与 25% 分位数的差值。下同。

图 3　GPM 降水在不同海拔区月降水的性能评估

Fig. 3　Performance evaluation of GPM satellite for monthly precipitation at different elevation zones

3.2　不同降水融合方法的精度评价

图 4 为年尺度和季节尺度各融合模型(卫星融合源为 GPMv07)精度评价指标(CC、BIAS、RMSE 和 NSE)的箱线图。从相关系数来看,各类数据均表现出冬季相关系数低于其他季节及年尺度的现象。在冬季以外的季节,各方法得到的融合降水的 CC 平均值略低于 GPMv07,但在冬季多数方法得到的融合降水的 CC 平均值均高于 GPMv07,其中随机森林法提升幅度最多

(0.11),其次为XGBoost方法。从相对误差来看,GPMv07在不同时间尺度的相对误差绝对值均高于0.20,且高估春夏秋季的降水,而低估冬季降水。各方法得到的融合降水相对误差的季节平均值均减小到0.20以内。均方根误差均呈现出年尺度值高于季节值,夏季均方根误差高于其他季节的特征,主要是由于RMSE的表现受降水量大小的影响,研究区夏季降水量多,则RMSE也表现为较大。不同时间尺度下各方法得到的融合降水均方根误差较GPMv07减小,减幅最高的为XGBoost方法,在春、夏、秋、冬季节分别减少3.52、15.98、3.07、1.55 mm。GPMv07在不同时间尺度下的纳什系数平均值为−1.38~0.10,说明GPMv07对年季尺度降水的再现能力较差。各方法得到的融合降水纳什系数均高于GPMv07,增幅最高的为XGBoost模型,在不同季节的增幅为0.35~1.74。总之,GPMv7虽然能较好地呈现多数季节降水的变化趋势,但高估冬季外其他季节的降水,且对冬季降水的变化趋势和平均值的再现能力低于其他季节,可能与冬季降雪难以模拟有关。各方法得到的融合季节降水的精度均高于GPMv7,其中XGBoost方法得到的融合降水精度最高。

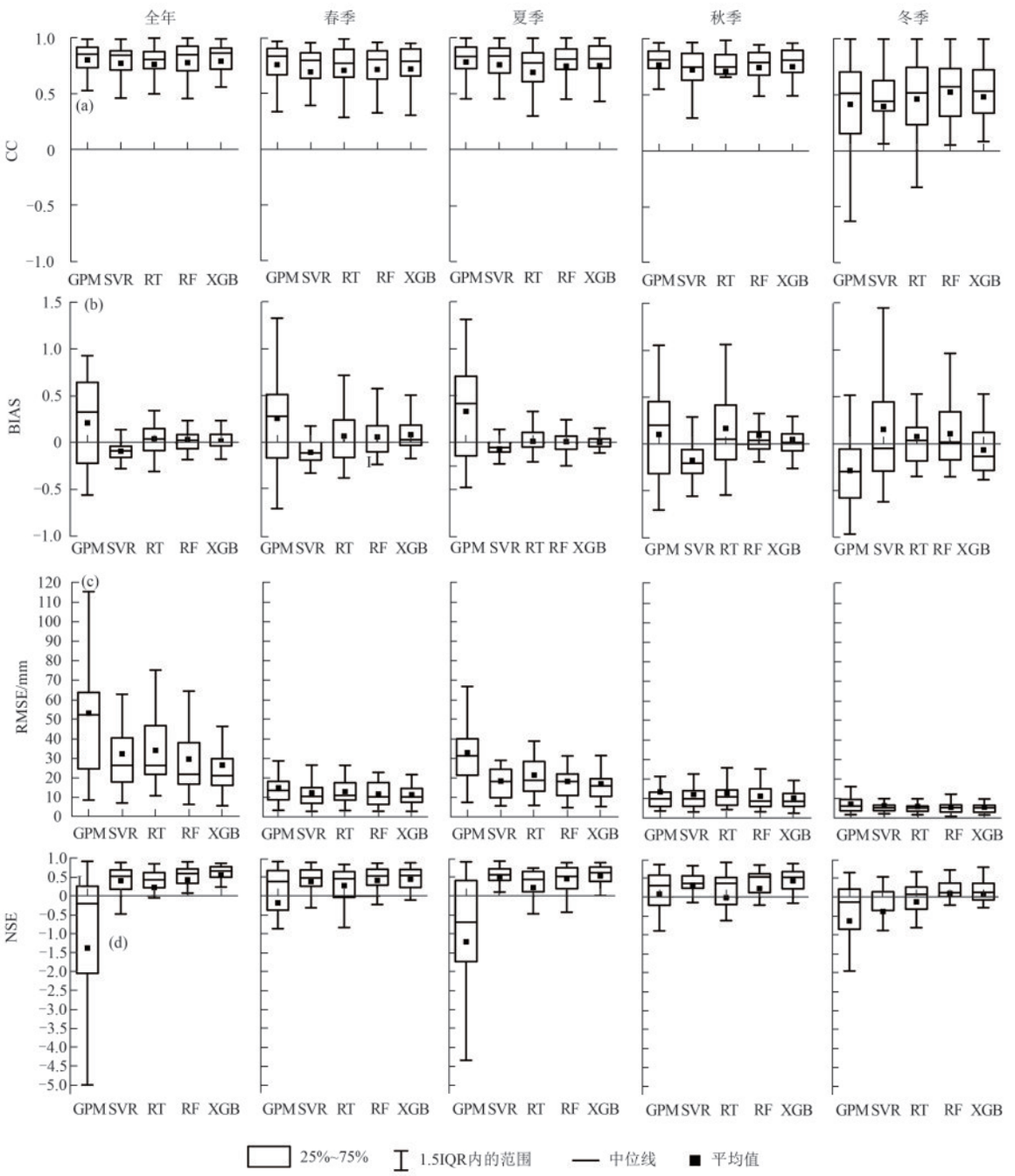
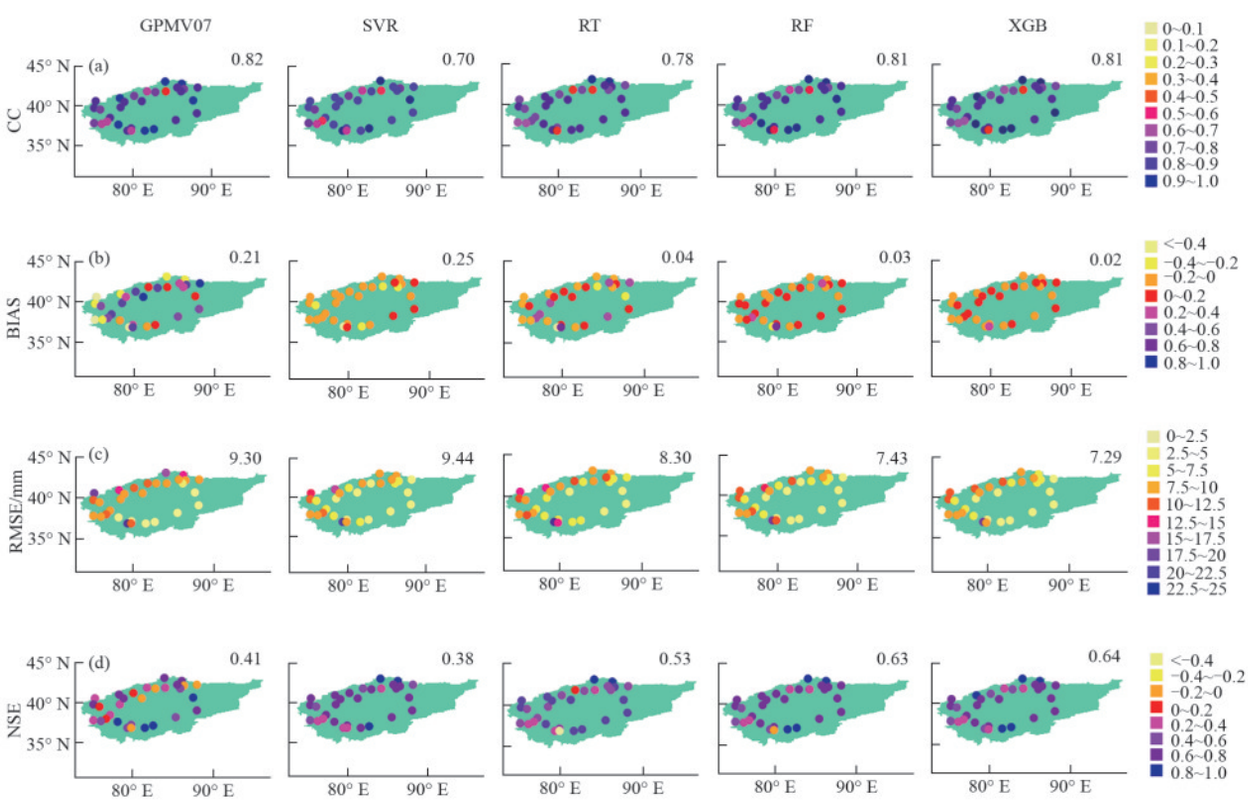


图4 GPMv07及各融合降水精度评价指标

Fig. 4 Accuracy evaluation indexes for GPMv07 and precipitation fusion methods

为了从不同角度验证各降水融合方法的有效性,本研究计算了各种方法模拟的融合降水在塔里木河流域30个站点的精度评价指标值。由图5可知,从相关系数的角度来看,GPMv07与多数站点月降水的相关系数在0.60以上,相关系数平均值为0.82,SVR、RF及XGBoost方法计算的融合降水的相关系数平均值与GPMv07相近。此外,GPMv07和不同方法计算的融合降水均呈现东南部站点的降水相关系数(>0.8)高于其他区域的特征。从相对误差来看,GPMv07高估多数区域的降水量,月降水平均误差为0.21。仅在20%站点处的相对误差绝对值低于0.20,在北部个别站点的月降水相对误差甚至高于0.80。与GPMv07相比,不同融合方法均显著降低月降水误差,其中精度最高的模型为XGBoost,其在97%站点处的相对误差绝对值低于0.20,站点平均相对误差为0.02;精度最低的融合模型为RT,其在70%站点处的相对误差绝对值在0.20以内,站点平均误差为0.04。从均方根误差来看,GPMv07的均方根误差平均值为9.30 mm,仅在27%的站点处的均方根误差低于7.5 mm,在西部和北部站点的均方根误差较大。不同方法得出的融合降水误差较GPMv07有所减小,其中均方根误差平均值最小的模型为XGBoost,比GPMv07均方根误差减小2.01 mm,且均方根误差低于7.50 mm的站点占比升高至53%,RT模型计算的融合降水的均方根误差平均值为8.30 mm,高于其他方法得到的融合降水的均方根误差。从纳什效率系数来看,GPMv07指标平均值为0.41,仅在37%的站点处月降水纳什系数不低于0.6,指标较低区位于西部和北部。XGBoost方法的精度最高,指标平均值较GPMv07提升约0.23,站点纳什系数不低于0.60的站点占比提升至70%。其次为RF法,精度最低的方法为RT法,指标平均值较GPMv07仅提升0.10。总之,GPMv07高估多数区域的月降水量,尤其在西部和北部站点的均方根误差较大,甚至部分站点难以较好地呈现月降水的变化趋势。XGBoost方法得到的融合月降水精度最高,与GPMv07相比,其相对误差绝对值在0.20以内的站点占比提升约77%,均方根误差减小2.01 mm,站点纳什系数不低于0.60的站点占比及纳什系数均值分别提升33%和0.23,而RT法得到的融合月降水精度最低。



注:图中每列包含4个子图,分别表示GPMv07和不同降水融合方法计算的月降水在CC、RMSE、BIAS和NSE指标方面的表现,每图的右上角数值为研究区所有站点相应统计指标的平均值。

图5　GPMv07及各融合降水精度评价指标的空间分布

Fig. 5 Spatial distribution of accuracy evaluation indexes for GPMv07 and precipitation fusion methods

3.3 不同卫星降水数据集对降水融合模型精度的影响

为评估不同卫星降水数据集作为融合源对融合模型精度的影响，本文分别采用GPMv07和GPMv06作为卫星降水融合源构建XGBoost模型，不同融合模型评价结果见图6。与GPMv06相比，基于GPMv07的融合降水能更好地再现年降水的变化趋势及平均值，相关系数和纳什效率系数均高于0.91，相对误差约0.02。相似地，在多数季节基于GPMv07的融合降水较GPMv06的融合降水的精度略有提升，相关系数提高0.01～0.02，均方根误差减小0.59～2.34 mm，纳什效率系数提升幅度为0.03～0.04。而在冬季，基于GPMv07和GPMv06的融合降水的精度相近。考虑到GPMv07的冬季降水精度显著高于GPMv06（表1），纳什效率系数甚至提高0.58，而基于二者的融合降水的精度相近，可推断出冬季降水融合模型精度对卫星降水数据误差的敏感性很小。而对于夏秋季节，GPMv07降水的纳什效率系数比GPMv06提升0.10～0.15，而基于不同版本GPM的融合降水的纳什效率系数差别仅为0.03，进一步说明卫星降水输入误差对融合模型XGBoost的精度影响较小。与TYSON等[26]的研究结果是一致，TYSON等[26]关注输入的不确定性对基于数据驱动模型的径流模拟精度的影响，结果表明，尽管输入数据精度有较大差异，但模拟径流的精度相近。

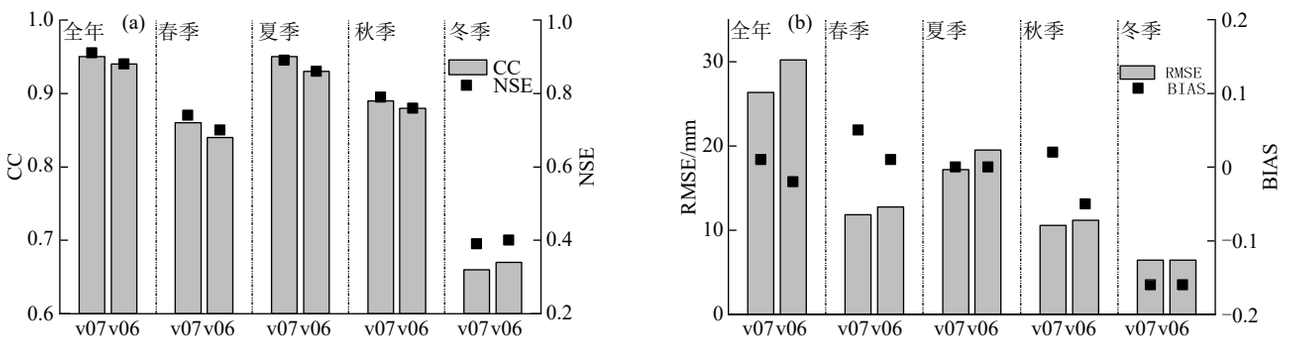


图6 基于不同卫星降水融合源构建的XGboost融合模型的性能评估

Fig. 6 Performance evaluation of XGboost fusion models based on different satellite precipitation products

3.4 考虑NDVI对降水响应的滞后性对融合模型精度的影响

已有研究[14]表明，NDVI可能对降水有3个月的滞后性。为验证NDVI滞后性是否对降水融合方法的精度产生影响，文中设置NDVI无滞后、滞后1～4月共5种情景，分别计算不同情景下XGBoost融合方法的精度。由图7可知，无论以GPMv07还是GPMv06作为卫星降水融合源时，考虑NDVI滞后3月或4月的融合方法在春夏季节的部分月份呈现更高的精度，如对于以GPMv06作为卫星降水融合源的XGBoost方法，考虑NDVI滞后3个月情景下的融合降水精度在4月、6—8月高于其他情景下的降水，与未考虑NDVI滞后的情景相比，其纳什效率系数升高0.03，均方根误差减小0.16～0.81 mm。而在秋冬季节月份（2月除外），无NDVI滞后情景下的融合降水精度最高，可见考虑NDVI的滞后对秋冬季节融合降水精度的影响很小。

4 讨 论

最新一代GPM降水数据产品（V07）于2023—2024年陆续发布，已有研究[27]关注GPMv06和GPM v07产品间的精度差异，Aksu评估GPMv07在土耳其的降水精度，指出GPM v07较GPM v06的降水精度有显著改进，但仍然存在低估高海拔降水的不足，与本研究的结论类似。本文对比2个版本GPM降水产品在不同季节的精度差异发现，GPMv07冬季的降水精度显著高于GPMv06，尤其在冬季纳什效率系数提高0.58，冬季降水精度的提升归因于GPMv07中算法的升级[28]，如GPROF算法的改进有助于增强对降雪事件的追踪，且地理位置的偏移也被修正。分析发现，春夏季节的部分月份考虑NDVI对降水的滞后性对融合模型的精度提升是正效应，而考虑NDVI对降水的滞后性对秋冬季节多数月份融合降水精度的影响较小，原因可能是积雪覆盖、秋冬季的植被落叶特征影响NDVI与降水之间的关系[14]。

本文对比4种机器学习方法（随机森林、支持向量机、XGBoost及回归树）对区域卫星降水校正模拟的效果差异发现，XGBoost方法表现出最高的精度，表明XGBoost方法能更好捕捉地面降水、卫星降水及其他辅助变量间的非线性关系。这主要归因于模型自身引入稀疏感知和正则化算法，通过约束树结构，降低模型方差[29]。为进一步验证XGBoost方法（GPMv07作为卫星降水融合源）得到的融合降水数

据的优势,本研究将其与可公开下载的GPM卫星观测融合产品AIMERG[30](GPM数据与APHRODITE数据的融合产品)进行对比,评估结果见图8。在月尺度及春冬季节,XGBoost方法融合模型计算的降水相关性略低于AIMERG数据。但XGBoost方法得到的融合降水在季月尺度下较AIMERG均表现出更高的纳什效率系数,月尺度下的效率系数增幅为0.44,季节尺度下效率系数增幅为0.97～3.57,其中夏季效率系数增幅最大。由误差指标(均方根误差和相对误差)精度对比可知,XGBoost方法得出的融合降水在月季尺度下均呈现更小的相对误差(－0.02～0.07),而AIMERG数据严重低估多数站点冬季降水量(BIAS＝－0.45),而高估其他季节的降水量(BIAS为0.46～0.60)。与AIMERG数据相比,XGBoost方法得出的融合降水在月尺度下的相对误差减小0.44,季节尺度下相对误差减小幅度为0.40～0.59。总之,XGBoost方法得到的融合降水在月和季节尺度下有更大的优势。

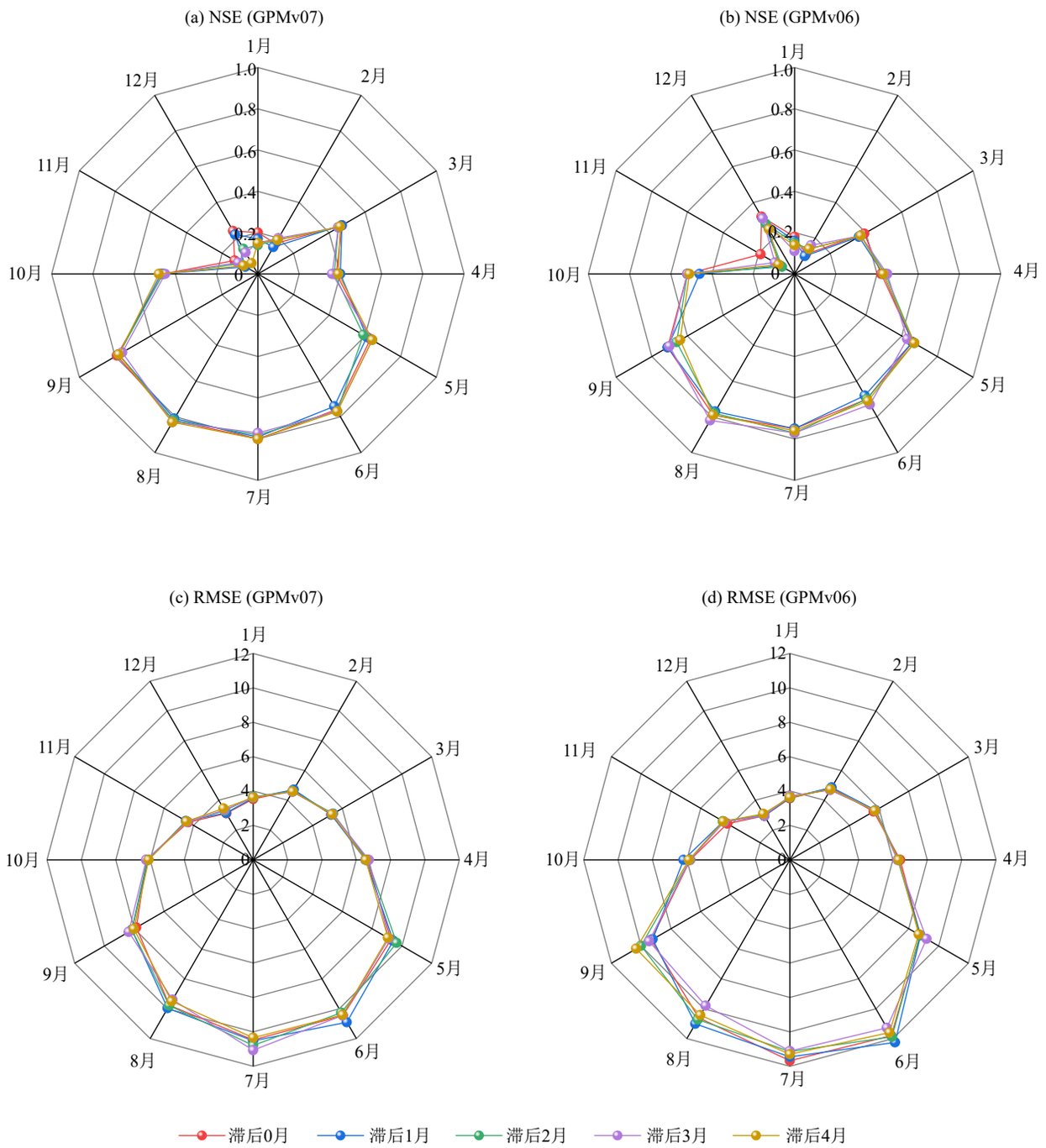


图7　考虑NDVI对降水滞后性的不同情景下融合降水精度评估

Fig. 7　Accuracy evaluation of precipitation fusion methods under different scenarios incorporating lagged response of NDVI to precipitation

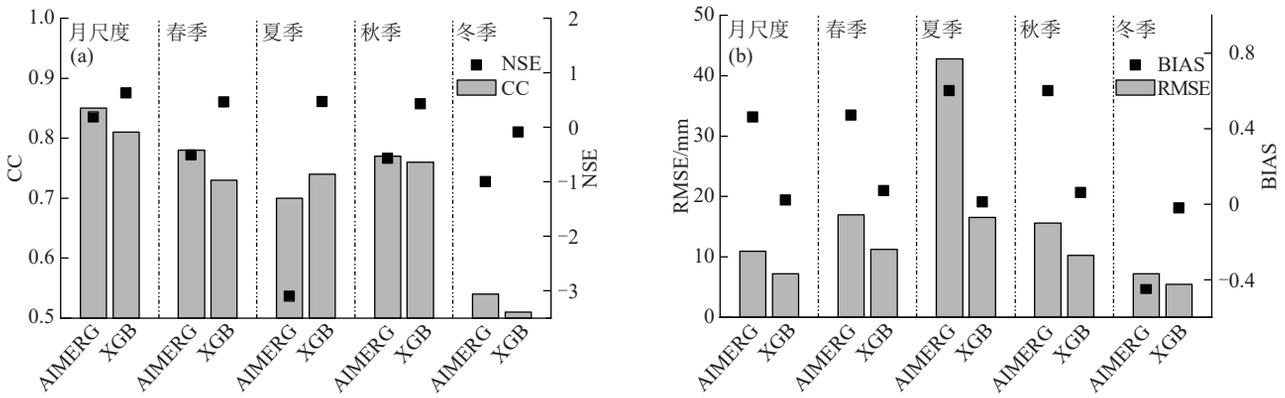


图 8 XGBoost 方法得到的融合降水与 AIMERG 降水产品的精度对比

Fig. 8 Accuracy comparison between fused precipitation obtained by XGBoost model and AIMERG precipitation product

5 结 论

1)GPMv07 在夏季和冬季的降水精度显著高于 GPMv06,尤其在冬季纳什效率系数提高 0.58,但冬季降水仍存在低估现象,可能与冬季降雪难以模拟有关。

2)在融合模型的评估中,XGBoost 方法表现出最高的精度,其在月尺度下的相对误差绝对值在 20% 以内的站点占比提升约 77%,纳什系数不低于 0.6 的站点占比及纳什系数平均值分别提升 33% 和 0.23,显著优于其他融合方法。

3)在春夏季节的部分月份考虑 NDVI 对降水的滞后性有利于提升 XGBoost 融合模型的精度,而考虑 NDVI 对降水的滞后性对秋冬季节多数月份融合降水精度的影响较小。

本研究成果丰富了卫星降水融合校正方法的研究,还可为区域水资源管理及水土流失预防研究提供数据参考。

参考文献:

[1] 姜大膀,王娜.IPCC AR6 报告解读:水循环变化[J].气候变化研究进展,2021,17(6):699-704.
JIANG D B, WANG N. Water cycle changes: Interpretation of IPCC AR6[J].Climate Change Research,2021,17(6):699-704.

[2] 申豪勇,李佳,王志恒,等.黄河支流汾河流域水资源开发利用现状及生态环境问题[J].中国地质,2022,49(4):1127-1138.
SHEN H Y, LI J, WANG Z H, et al. Water resources utilization and eco-environment problem of Fenhe River, branch of Yellow River[J].Geology in China, 2022, 49(4):1127-1138.

[3] 董甲平,冶运涛,顾晶晶,等.滦河流域遥感降水降尺度多时间特性分析[J].水力发电学报,2022,41(8):77-91.
DONG J P, YE Y T, GU J J, et al. Multi-temporal characterization analysis of remotely sensed precipitation downscaling in the Luanhe River basin, China[J].Journal of Hydroelectric Engineering,2022,41(8):77-91.

[4] 熊景华,郭靖,郭生练,等.基于多源降水数据估算澜湄流域可能最大降水[J].水力发电学报,2022,41(9):77-86.
XIONG J H, GUO J, GUO S L, et al. Estimating probable maximum precipitation based on multisource data of precipitation in the Lancang-Mekong River basin[J].Journal of Hydroelectric Engineering,2022,41(9):77-86.

[5] 吴炳方,朱伟伟,曾红伟,等.流域遥感:内涵与挑战[J].水科学进展,2020,31(5):654-673.
WU B F, ZHU W W, ZENG H W, et al. Watershed remote sensing: Definition and prospective[J].Advances in Water Science,2020,31(5):654-673.

[6] 王忠静,石羽佳,张腾.TRMM 遥感降水低估还是高估中国大陆地区的降水?[J].地球科学进展,2021,36(6):604-615.
WANG Z J, SHI Y J, ZHANG T. Does TRMM precipitation underestimate or overestimate in China's mainland? [J]. Advances in Earth Science, 2021, 36(6):604-615.

[7] 南天一,陈杰,丁智威,等.基于深度学习的青藏高原多源降水融合[J].中国科学地球科学,2023,53(4):836-855.
NAN T Y, CHEN J, DING Z W, et al. Deep learning-based multi-source precipitation merging for the Tibetan Plateau[J].Science China Earth Sciences,2023,66(4):852-870.

[8] 余辉,梁镇涛,鄢宇晨.多来源多模态数据融合与集成研究进展[J].情报理论与实践,2020,43(11):169-178.
YU H, LIANG Z T, YAN Y C. Review on multi-source and multi-modal data fusion and integration[J]. Information Studies (Theory and Application),2020,43(11):169-178.

[9] 卢新玉,刘艳,王秀琴,等.新疆地区多源降水融合试验

[J].干旱区研究,2020,37(5): 1223-1232.
LU X Y, LIU Y, Wang X Q, et al. Multisource precipitation data merging experiment in Xinjiang [J]. Arid Zone Research,2020,37(5): 1223-1232.

[10] 潘旸,沈艳,宇婧婧,等.基于贝叶斯融合方法的高分辨率地面-卫星-雷达三源降水融合试验[J].气象学报,2015,73(1): 177-168.
PAN Y, SHEN Y, YU J J, et al. An experiment of high-resolution gauge-radar-satellite combined precipitation retrieval based on the Bayesian merging method[J]. Acta Meteor Sinica,2015, 73(1):177-186.

[11] 石羽佳,王忠静,索滢.基于多源数据融合的海河流域降水资源评价[J].水科学进展,2022,33(4):602-613.
SHI Y J, WANG Z J, SUO Y. Evaluation of Haihe River basin precipitation resources based on multisource data fusion[J].Advances in Water Science,2022,33(4): 602-613.

[12] 阮惠华,张钧民,许剑辉,等.考虑降水时间相关性的地面观测-雷达-卫星遥感逐时降水融合方法研究[J].热带气象学报,2023,39(3):300-312.
RUAN H H, ZHANG J M, XU J H, et al. An XGBoost-based geostatistical data fusion method for integrating hourly gauge-radar-satellite precipitation data by considering the temporal correlation characteristics of precipitation[J]. Journal of Tropical Meteorology, 2023, 39(3):300-312.

[13] LU X Y, LI J, LIU Y, et al. Quantitative precipitation estimation in the Tianshan Mountains based on machine learning[J].Remote Sensing,2023,15(16):e3962.

[14] KARBALAYE GHORBANPOUR A, HESSELS T, MOGHIM S, et al. Comparison and assessment of spatial downscaling methods for enhancing the accuracy of satellite-based precipitation over Lake Urmia basin[J]. Journal of Hydrology,2021,596:e126055.

[15] 邓文彬,侯雪晴.基于机器学习算法构建新疆积雪覆盖率预测模型[J].应用基础与工程科学学报,2024,32(6):1664-1677.
DENG W B, HOU X Q. Construction of a snow cover prediction model in Xinjiang based on machine learning algorithm[J]. Journal of Basic Science and Engineering, 2024,32(6):1664-1677.

[16] 孙倩,阿丽亚·拜都热拉,依力亚斯江·努尔麦麦提.归一化植被指数对陆地水储量和降水变化的响应研究:以塔里木河流域为例[J].中国农村水利水电,2018(2): 54-59.
SUN Q, Aliya·Baidourela, Ilyas·Nurmuhammat. The response of NDVI to the changes of terrestrial water storage and precipitation in Tarim basin [J]. China Rural Water and Hydropower,2018(2):54-59.

[17] 李文文.气候变化与人类活动影响下的塔里木河流域水土资源利用研究[D].南京:南京信息工程大学,2022.
LI W W. Study on the utilization of water and soil resources in Tarim River basin under the influence of climate change and human activities[D]. Nanjing: Nanjing University of Information Science and Technology,2022.

[18] 李媛媛,宁少尉,丁伟,等.最新GPM降水数据在黄河流域的精度评估[J].国土资源遥感,2019,31(1):164-170.
LI Y Y, NING S W, DING W, et al. The evaluation of latest GPM-Era precipitation data in Yellow River basin [J]. Remote Sensing for Land and Resources, 2019, 31(1):164-170.

[19] 顾晶晶,冶运涛,董甲平,等.滦河流域遥感反演降水产品高精度空间降尺度方法[J].南水北调与水利科技(中英文),2021,19(5):862-873.
GU J J, YE Y T, DONG J P, et al. A high-precision spatial downscaling method for remotely sensed precipitation data in the Luanhe River basin[J]. South-to-North Water Transfers and Water Science and Technology, 2021,19(5):862-873.

[20] 蔡超,冉晓婷,薛伟,等.大数据背景下分布式支持向量回归模型研究[J].系统科学与数学,2023,43(4):1081-1092.
CAI C, RAN X T, XUE W, et al. Research on distributed support vector regression model under the background of big data[J]. Journal of Systems Science and Mathematical Sciences,2023,43(4):1081-1092.

[21] 王如冰,蔡喜运.基于梯度提升回归树的有机污染物生物-沉积物积累因子预测模型[J].生态毒理学报,2023,18(4):22-33.
WANG R B, CAI X Y. Biota-sediment accumulation factor models of organic chemicals in benthic invertebrates with gradient boosting regression tree[J]. Asian Journal of Ecotoxicology,2023,18(4):22-33.

[22] 邓玉睿,程旭东,唐芳,等.基于多元线性回归分析和随机森林算法的水稻贮藏霉变风险控制[J].中国科学技术大学学报,2022,52(1):44-51.
DENG Y R, CHENG X D, TANG F, et al. The control of moldy risk during rice storage based on multivariate linear regression analysis and random forest algorithm [J].Justc,2022,52(1):44-51.

[23] 谢永强.集成多源数据与XGBoost算法京津冀地区土壤水分空间反演[J].地理空间信息,2024,22(12):20-24.
XIE Y Q. Spatial inversion of soil moisture in the Beijing-Tianjin-Hebei Region using integrated multi-source data and XGBoost algorithm [J]. Geospatial Information, 2024,22(12):20-24.

[24] 谭洁,危千骏,廖朝阳,等.基于XGBoost-SHAP可解释机器学习模型的城市形态与地表温度的关系[J].应用生态学报,2025,36(3):659-670.
TAN J, WEI Q J, LIAO Z Y, et al. Relationship between urban form and surface temperature based on

XGBoost SHAP interpretable machine learning model [J]. Chinese Journal of Applied Ecology, 2025, 36(3): 659-670.

[25] 纪昌明,赵亚威,张验科,等.考虑弃水电量成本的短期电力电量平衡模型[J].水力发电学报,2021,40(3):50-63.
JI C M, ZHAO Y W, ZHANG Y K, et al. Short-term power balance model considering cost of abandoned hydropower [J]. Journal of Hydroelectric Engineering, 2021,40(3):50-63.

[26] TYSON C, LONGYANG Q Q, NEILSON B T, et al. Effects of meteorological forcing uncertainty on high-resolution snow modeling and streamflow prediction in a mountainous karst watershed [J]. Journal of Hydrology, 2023,619:e129304.

[27] AKSU H, YALDIZ S G. Performance comparison of GPM IMERG V07 with its predecessor V06 and its application in extreme precipitation clustering over Türkiye[J].Atmospheric Research,2025,315:e107840.

[28] WANG Y J, LI Z, GAO L, et al. Comparison of GPM IMERG version 06 final Run products and its latest version 07 precipitation products across scales: Similarities, differences and improvements[J].Remote Sensing,2023, 15(23):e5622.

[29] 廉睿.XGBoost算法在四川省GPM降水数据降尺度中的应用[J].水电能源科学,2021,39(10):14-17.
Lian R L. Application of XGBoost algorithm in downscaling of GPM precipitation data in Sichuan Province [J]. Water Resources and Power,2021,39(10):14-17.

[30] MA Z Q, XU J T, ZHU S Y, et al. AIMERG: A new Asian precipitation dataset (0.1°/half-hourly, 2000-2015) by calibrating GPM IMERG at daily scale using APHRODITE [J]. Earth System Science Data, 2020: 1525-1544.

(上接第418页)

[34] 徐万里,刘骅,张云舒,等.施肥深度、灌水条件和氨挥发监测方法对氮肥氨挥发特征的影响[J].新疆农业科学,2011,48(1):86-93.
XU W L, LIU H, ZHANG Y S, et al. Effects of fertilization depth, irrigation conditions, and ammonia volatilization monitoring methods on ammonia volatilization characteristics of nitrogen fertilizers[J].Xinjiang Agricultural Sciences,48(1):86-93.

[35] 杨淑莉,朱安宁,张佳宝,等.不同施氮量和施氮方式下田间氨挥发损失及其影响因素[J].干旱区研究,2010,27(3):415-421.
YANG S L, ZHU A N, ZHANG J B, et al. Field ammonia volatilization losses and influencing factors under different nitrogen application rates and methods[J]. Arid Zone Research,2010,27(3):415-421.

[36] 范思思.水分状况对黑土硝化反硝化功能基因丰度及N_2O排放的影响[D].辽宁 大连:大连交通大学,2018.
FAN S S. Effects of water conditions on the abundance of nitrification and denitrification functional genes and N_2O emissions in black soil[D].Dalian, Liaoning: Dalian Jiaotong University,2018.

[37] 白芳芳,乔冬梅,李平,等.减氮对麦玉轮作农田土壤反硝化细菌群落结构和多样性的影响[J].农业环境科学学报,2024,43(6):1338-1349.
BAI F F, QIAO D M, LI P, et al. Effects of nitrogen reduction on denitrifying bacterial community structure and diversity in wheat-maize rotation farmland soil [J]. Journal of Agro-Environment Science,43(6):1338-1349.

[38] 汪思佳,王春霞,张景瑞,等.减氮条件下基于AquaCrop模型的北疆膜下滴灌棉花水氮制度优化[J].水土保持学报,2024,38(3):314-324.
WANG S J, WANG C X, ZHANG J R, et al. Optimization of water and nitrogen regimes for drip-irrigated cotton under mulch in northern Xinjiang based on AquaCrop model under nitrogen reduction conditions[J].Journal of Soil and Water Conservation,38(3):314-324.